Université de Montréal

Modèle bayésien pour les prêts investisseurs

par

Mathieu Bouvrette

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en Statistique

octobre 2006

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Modèle bayésien pour les prêts investisseurs

présenté par

Mathieu Bouvrette

a été évalué par un jury composé des personnes suivantes :

Louis Doray
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Daniel Pouliot
(co-directeur)

Manuel Morales
(membre du jury)

Mémoire accepté le:
26 octobre 2006

SOMMAIRE

Les arbres de décision ont été développés en 1984 par Breiman, Freidman, Olshen et Stone et ils sont composés de deux parties, soient l'arbre de classification et l'arbre de régression. De plus, ces modèles de prévision sont simples, faciles à interpréter et ils supportent les variables explicatives nominales et catégorielles, contrairement à d'autres méthodes qui n'utilisent qu'un des deux types.

Ainsi, dans le cadre de la bourse d'études supérieures à incidence industrielle du CRSNG, nous élaborons sur les arbres de classification. Nous présentons aussi une approche bayésienne des arbres de classification discutée dans les articles de Chipman H., George E. et McCullock R. (2005), Schetinin (2004) et Denison D., Malick B. et Smith A. (1998).

De plus, nous présentons plusieurs autres modèles de prévision, dont la régression logistique, la régression « probit » et l'analyse discriminante afin de comparer avec l'approche bayésienne d'un modèle d'arbre de classification. Ces modèles sont comparés sur leur performance de prévision du risque de mauvais payeur (probability of default) pour les prêts investisseurs (prêts avec garantie sur fonds mutuels) de la Banque Nationale du Canada.

Mots clés : Modèle bayésien, Défaut de paiement, Prêts investisseurs, Fonds mutuels, Arbre de classification, Forêt d'arbres, Arbre consensus, Analyse discriminante, Régression logistique/probit, Chaînes de Markov à sauts réversibles, Test t, Test de Wilcoxon, d Cohen, Mesures d'association.

SUMMARY

Decision trees were developed in 1984 by Breiman, Freidman, Olsen and Stone and are made up of two parts, the classification tree and the regression tree. In addition, these anticipation models are simple, easy to interpret and they support the explanatory nominal and categorical variables, contrary to other methods that only utilize one of these two types.

Thus, within the context of the CRSNG grant for higher education in the industrial field, we elaborate on classification trees. We also present a Bayesian approach of the classification trees which are discussed in the articles by Chipman H., George E. and McCullock R. (2005), Schetinin (2004) and Denison D., Malick B. and Smith A. (1998)

Furthermore, we present other anticipation models such as logistic regression, the « probit » regression and a discriminatory analysis, in order to compare the Bayesian approach with another classification tree models. These models are compared by the anticipated performance of the risk involved for a bad payer (probability of default) for investment loans (loan with a guarantee on mutual funds) from the National Bank of Canada.

Keywords : Bayesian model, Default of payment, Investment loans, Mutual funds, Classification tree, Forest of trees, Consensus tree, Discriminatory analysis, Logistic/probit regression, Reversible jump Markov chain, T test, Wilcoxon test, d Cohen, Association measures.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

J'aimerais remercier mon directeur de recherche Jean-François Angers pour tout le temps et les efforts qu'il m'a consacré pendant ces dernières années. J'aimerais aussi remercier Jean-François pour ses conseils et tout le cheminement scolaire qu'il m'a permis de parcourir.

Je voudrais remercier Chantal Legault et la Banque Nationale du Canada pour avoir accepté de se lancer dans ce projet, de l'avoir financé et d'y avoir cru jusqu'au bout. Aussi, je voudrais remercier mes trois directeurs de la Banque Nationale, Chantal Legault, Yann Jodoin et Daniel Pouliot, pour leur temps, leurs idées et leur support.

De plus, j'aimerais remercier le conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour sa contribution financière importante en m'octroyant une bourse d'études supérieures à incidence industrielle. Cette aide financière m'a permis d'obtenir une expérience de recherche intéressante à la Banque Nationale tout en poursuivant mes études aux cycles supérieures.

Finalement, j'aimerais remercier ma copine et toute ma famille pour leur support moral, financier et leurs encouragements.

INTRODUCTION

Ce mémoire est le fruit d'un projet commun entre le Département de mathématiques et de statistique de l'Université de Montréal et la Banque Nationale du Canada dans le cadre du programme de bourses d'études supérieures à incidence industrielle du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG). Ce programme permet à des étudiants du deuxième ou troisième cycle de faire de la recherche dirigée et d'appliquer les résultats obtenus dans le milieu du travail. Ainsi, le sujet du mémoire vient d'un problème du secteur du risque aux particuliers, soit le manque d'information pour une prévision adéquate du risque.

La création d'un nouveau type de prêt engendre parfois des complications pour la gestion du risque, car l'absence d'historique empêche de faire une analyse statistique avec une bonne puissance. Pour pallier ce problème, nous proposons l'utilisation de la statistique bayésienne. Donc, en connaissant le comportement de la délinquance de prêts similaires, nous pouvons récolter de l'information utile à la prévision du risque de ce nouveau type de prêt. Bref, nous allons tenter d'augmenter la performance des prévisions par des modèles bayésiens.

Dans ce mémoire, nous nous intéressons à la prévision du retard sur le remboursement d'un prêt investissement de la Banque Nationale. Ce type de prêt est d'autant plus intéressant qu'il présente une double problématique. En plus d'un historique limité, celui-ci est utilisé comme levier financier et, par conséquent, les fonds choisis peuvent influencer grandement une délinquance éventuelle du client. Ainsi, nous allons appliquer les méthodes statistiques standards et de nouvelles

méthodes pour résoudre ces problématiques.

Plusieurs méthodes statistiques peuvent être utilisées dans la prévision du risque. Par exemple, nous pouvons employer une analyse par région décrite dans le livre de Spath (1980) ou bien une analyse discriminante énoncée dans Bishop, Fienberg, et Holland (1975). Hosmer et Lemeshow (2000) propose une régression logistique tandis que Finney (1971) décrit plutôt une régression logistique avec un modèle « *probit* ». De plus, Breiman, Friedman, Oslhen et Stone (1984) ont popularisé les arbres de classification et de régression ou « *Classification And Regression Tree* (CART) » comme nouvelle méthode statistique et elle a été reprise, plus tard, par plusieurs autres, dont Quinlan (1986) et Timofeev (2005). Finalement, certains auteurs ont mis de l'avant les arbres de classification bayésiens, par exemple Chipman et McCulloch (1998), Denison, Mallick et Smith (1998), Schetinin (2004) et Denison, Holmes, Malick et Smith (2002).

Tout d'abord, dans le premier chapitre, nous énonçons les différentes définitions utilisées tout au cours du mémoire. Ensuite, nous décrivons le cheminement qui nous a permis de choisir la méthode statistique la plus apropriée pour notre situation avec un survol des méthodes usuels de prédiction. Finalement, nous abordons l'arbre de décision et les avantages et inconvénients de l'utilisation de cette méthode.

Dans le deuxième chapitre, nous présentons les deux bases de données utilisées pour l'application des modèles. Nous élaborons un aperçu général de ces données, de leur distribution et des différents liens entre elles. De plus, nous présentons et comparons deux méthodes d'imputations des valeurs manquantes et nous justifions leur application. Ensuite, nous effectuons une analyse du comportement des fonds reliée au prêt détenu. En dernier lieu, la relation de dépendance entre la valeur du fonds et la présence de délinquance sur le remboursement d'un prêt est illustrée, et ce pour un prêt investisseur ou personnel.

Dans le troisième chapitre, l'arbre de classification de Breiman, Friedman, Oshlen et Stone (1984) est présenté en détail avec les règles de séparation les plus fréquemment utilisées. De plus, un l'algorithme itératif sur la minimisation de l'impureté est illustré et deux méthodes pour choisir le « meilleur » arbre sont énoncées. Finalement, nous terminons ce chapitre par la démarche à suivre en présence de valeurs manquantes.

Dans le quatrième chapitre, nous construisons l'arbre de classification avec une approche bayésienne avec la méthode énoncée par Denison, Mallick et Smith (1998). Ensuite, le choix des fonctions de densité *a priori* pour les différentes parties d'un arbre de classification, présenté par Chipman, George et McCulloch (1998), sont énoncées. De plus, nous abordons d'autres choix de fonctions *a priori* qui peuvent s'appliquer et donner des résultats intéressants. En dernier lieu, l'algorithme de la méthode de Monte Carlo markovienne à sauts réversibles, énoncé par Green (1995) et appliqué par Schetinin (2004), est expliqué.

Finalement, dans le cinquième chapitre, nous présentons une méthode dite « bagging » ou par vote permettant de construire une forêt d'arbres bayésiens introduit par Breiman (1996). En raison de la complexité des forêts, nous présentons également un arbre consensus comme solution de rechange. Dans un dernier temps, nous énonçons les différentes mesures d'association utilisées pour quantifier la prédiction et nous dévoilons les résultats obtenus pour tous les modèles présentés dans ce mémoire.

Chapitre 1

PRÉLIMINAIRES

Dans ce chapitre, nous commençons par définir les objectifs du mémoire et nous survolons différents concepts théoriques nécessaires ou utiles à la compréhension des chapitres à venir. Ensuite, nous présentons quelques tests statistiques. Finalement, nous passons au travers de quatre modèles de prévision et nous regardons brièvement leur algorithme afin de pouvoir déterminer lequel nous donne la meilleure performance sur notre échantillon.

1.1. OBJECTIFS

Les objectifs de ce mémoire sont les suivants :

1- analyser l'impact du comportement du fonds en garanti sur la délinquance du remboursement d'un prêt investissement.

2- Déterminer la probabilité qu'un client soit délinquant dans les 4 prochains mois en utilisant les informations de son prêt sur les douze derniers mois.

1.2. DÉFINITIONS

Tout d'abord, voici quelques définitions standards utilisées tout au cours de ce mémoire.

Définition 1. Nous commençons par définir certains termes utilisés dans le domaine bancaire.

Défaut : *90 jours de retard et plus* ;
Délinquance : *client ayant un ou plusieurs retards* ;
Fenêtre : *relevé d'informations à un moment précis* ;
Retard : *non paiement de la somme due à la date de la facture* ;
Présence d'un retard : *entre 1 et 90 jours de retard.*

Aussi, nous pouvons définir le prêt investissement comme suit.

Définition 1.2.1 (Le prêt investissement). *Le prêt investissement est un prêt permettant de financer à taux réduit l'achat de fonds mutuels. En somme, ce type de prêt est utilisé à titre de levier financier dans le but d'accroître le montant investi, de diversifer son portefeuille, ou de compléter toute autre stratégie. Tous les fonds communs de placement acquis avec ce prêt sont détenus comme valeur de garantie et aucun dépôt initial n'est demandé. Ce type de prêt est très récent et nous avons peu d'historique sur la qualité du remboursement de celui-ci.*

De plus, l'étude d'un prêt peut se décomposer en deux parties, le moment de l'émission du prêt et le suivi du prêt. Les informations recueillies lors de l'émission du prêt sont plus difficiles à utiliser puisqu'elles peuvent être très anciennes et par conséquent, d'une moins grande fiabilité. Par contre, ces données s'avérent riches en informations diversifées. Par exemple, des renseignements sur la valeur de la propriété pour un prêt hypothécaire sont demandés au moment de l'émission, mais ceux-ci sont rarement mis à jour lors du suivi du prêt à cause du coût de collecte. Alors, nous avons conservé uniquement quelques informations à l'émission du prêt dans la construction de la base de données.

Une mesure intéressante dans les prêts investissements est l'écart entre le prêt et la valeur du fonds que nous appelons les capitaux propres. Pour mieux comprendre cette mesure, énonçons une définition formelle des capitaux propres.

Définition 1.2.2 (Les capitaux propres). *Le droit des propriétaires sur l'actif total correspond aux capitaux propres. Il est égal au total de l'actif moins le total du passif. Pour déterminer ce qui appartient aux propriétaires, il faut soustraire les créances, c'est-à-dire le passif (dettes et obligations) de l'actif (ressources disponibles), voir Weygandt, Kieso, Kimmel et Trenholm (2003).*

En continuant, nous définissons le type de statistique utilisée pour résoudre le problème d'absence d'historique, soit l'inférence bayésienne.

Définition 1.2.3 (L'inférence bayésienne). *L'inférence bayésienne est une méthode logique permettant de calculer ou réviser la probabilité d'une hypothèse. Cette méthode utilise des combinaisons de probabilités et le théorème de Bayes. Alors, un modèle bayésien est composé d'un modèle statistique paramétrique provenant des observations et d'une distribution a priori sur les paramètres inconnus.*

Par conséquent, l'inférence bayésienne consiste à tenir compte des informations que nous connaissons sur les prêts investissements avant d'entreprendre l'étude. Ainsi, l'inférence bayésienne fait le choix de modéliser les attentes en début de processus (quitte à réviser ce premier jugement après la première expérience). Celle-ci est préférable dans les cas où les données sont rares et/ou dispendieuses à obtenir. Advenant l'utilisation d'un grand nombre de données, les résultats de la statistique classique et de l'inférence bayésienne seraient asymptotiquement les mêmes.

1.3. TESTS STATISTIQUES

Tout au cours de ce document, nous avons besoin de plusieurs tests statistiques dont le test t pour échantillons indépendants, le test de Wilcoxon (1945) et le d de Cohen (1988).

1.3.1. Le test t pour échantillons indépendants

Le test t permet de calculer la probabilité que l'hypothèse nulle soit vérifiée. L'hypothèse nulle et l'hypothèse alternative pour le test t sont données par :

$$H_0 : \mu_1 = \mu_2, \text{ contre } H_a : \mu_1 \neq \mu_2,$$

où μ_1 et μ_2 sont les moyennes de la population 1 et 2 respectivement. Nous supposons que les deux populations de même variance suivent une loi normale. Ainsi, le test t permet de vérifier si deux groupes, provenant de populations normales, possèdent la même moyenne.

Tout d'abord, pour faire un test t, il faut vérifier l'égalité des variances avec l'équation :

$$\max(S_X, S_Y) < 2\min(S_X, S_Y)$$

où X et Y représentent les deux échantillons et S leur écart-type.

1.3.1.1. *Égalité des variances*

Nous calculons un estimé de la variance :

$$\widehat{\sigma}^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}.$$

Ensuite, à l'aide d'une table de la loi de Student, nous calculons la probabilité de ne pas rejeter H_0 lorsque celle-ci est supposée vraie (valeur-p) :

$$\text{valeur-p} = 2\Pr\left(T_{n_1+n_2-2} \geqslant |t_{obs}|\right),$$

où

$$t_{obs} = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} \text{ et } ddl = n_1 + n_2 - 2. \tag{1.3.1}$$

1.3.1.2. *Variances inégales*

Dans le cas de variances inégales, il n'existe pas de test exact, mais nous pouvons toujours utiliser l'approximaition suivante :

$$\text{valeur-p} = 2\Pr\left(T_{ddl} \geqslant |t_{obs}|\right),$$

où

$$t_{obs} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/n_1 + S_Y^2/n_2}} \text{ et } ddl = \frac{\left(S_X^2/n_1 + S_Y^2/n_2\right)^2}{\frac{(S_X^2/n_1)^2}{n_1-1} + \frac{(S_Y^2/n_2)^2}{n_2-1}}.$$

Pour plus de détails sur le test t, voir Goulden (1956). Par contre, nous ne pouvons faire ce test si l'hypothèse de la normalité n'est pas vérifiée. Alors, comme solution, nous employons le test de la somme des rangs de Wilcoxon.

1.3.2. Le test de la somme des rangs de Wilcoxon

Le test de la somme des rangs de Wilcoxon calcule la même probabilité que pour le test t, c'est-à-dire la probabilité que l'hypothèse nulle ne soit pas rejetée. Aussi, ce test est dit non paramétrique, sans hypothèse sur la distribution des populations. De plus, le test de Wilcoxon s'utilise lorsque l'échantillon est de petite taille ou que les distributions des deux populations ne sont pas normales. De plus, il est basé sur la somme des rangs des données de chacune des classes. Ainsi, plus l'écart entre la somme des rangs des classes est élevé, plus il existe une différence entre les deux classes pour cette variable et vice-versa.

Pour faire un test de la somme des rangs de Wilcoxon, il suffit d'appliquer les étapes suivantes (voir Wilcoxon, 1945) :

(1) combiner les 2 échantillons

$$\begin{aligned} (z_1, \ldots, z_{n_1}) &= (x_1, \ldots, x_{n_1}), \\ (z_{n_1+1}, \ldots, z_{n_1+n_2}) &= (y_1, \ldots, y_{n_2}), \end{aligned}$$

où $(z_1, \ldots, z_{n_1+n_2})$ est l'échantillon combiné et n_1 et n_2 représentent les tailles des deux échantillons.

(2) Ordonner les observations des deux échantillons combinés,

$$z_{(1)} \leqslant z_{(2)} \leqslant \ldots \leqslant z_{(n_1+n_2)},$$

où $z_{(i)}$ représente la i^e observation ordonnée.

(3) Attribuer le rang à chacune des observations, $\vec{r} = (1, 2, \ldots, n_1 + n_2)$. Si deux ou plusieurs observations sont égales, remplacer leur rang par la moyenne de leur rang. Par exemple, si $x^{(11)} = x^{(12)}$, remplacer leur rang par 11,5.

(4) Additionner les rangs d'un groupe, $W_x = \sum_{i \in groupe} r_i$.

(5) Calculer $W_{obs} = \min(W_x, n_1(n_1 + n_2 + 1) - W_x)$.

(6) Rejeter l'hypothèse H_0 si $W_{obs} \leqslant W_{0,025}$, où la valeur de $W_{0,025}$ se trouve dans une table de rang de Wilcoxon.

Cette démarche résume le test de Wilcoxon.

Les deux tests précédents, soit le test t et le test de Wilcoxon, dépendent de la taille de l'échantillon. Donc, pour une petite taille d'échantillon, il se peut que des différences importantes ne soient pas significatives et, pour une grande taille d'échantillon, les très petites différences peuvent être très significatives. Ainsi, ces deux tests reflètent la grandeur de l'effet, c'est-à-dire la force de la relation entre les deux groupes, et la taille de l'échantillon. Dans ce mémoire, nous manipulons des grosses bases de données et nous voulons un test qui ne dépend que de la grandeur de l'effet, indépendamment de la taille de l'échantillon. Alors pour ce faire, nous introduisons la mesure du d de Cohen.

1.3.3. Le d de Cohen

Le d de Cohen (1988) se calcule différemment en fonction de la variance et de la taille de l'échantillon.

(1) Égalité des variances : $d = \frac{\bar{X}-\bar{Y}}{S}$ où S est l'écart-type d'un des deux groupes. (L'hypothèse de l'égalité des variances se calcule de la même façon que pour le test t.)

TABLEAU. 1.1. Interprétation du d de Cohen

Interprétation de Cohen	Grandeur effet (d)	% non chevauchement densité
Grand	$\geqslant 0,8$	$\geqslant 47,4$ %
Moyen	entre 0,3 et 0,8	entre 21,3 % et 47,4 %
Faible	$< 0,3$	$< 21,3$ %

(2) Inégalité des variances : $d = \frac{\bar{X}-\bar{Y}}{\sqrt{(S_X^2+S_Y^2)/2}}$.

(3) Échantillons de tailles égales : $d = \frac{2t_{obs}}{\sqrt{n-2}}$ où n est le nombre d'observations par groupes et t_{obs} provient de l'équation (1.3.1).

(4) Échantillons de tailles inégales : $d = \frac{t_{obs}(n_1+n_2)}{\sqrt{(N-2)n_1n_2}}$ où N est le nombre total d'observations et t_{obs} provient aussi de l'équation (1.3.1).

Si les deux hypothèses (taille ou variance) ne sont pas respectées simultanément, nous ne pouvons calculer le d de Cohen. Par contre, en choisissant la variance la plus faible des deux groupes, nous pouvons déterminer une borne inférieure au d de Cohen. Cette mesure s'interprète selon le tableau 1.1.

Par exemple, dans le tableau 1.1, si nous prenons un d de Cohen égal à 2, nous avons 81,1 % des données que nous pouvons séparer dans chacun des deux groupes et 18,9 % qui peuvent appartenir aussi bien à l'un qu'à l'autre.

1.4. DIFFÉRENTS MODÈLES STATISTIQUES

Un des objectifs de ce projet est de déterminer la probabilité qu'un client soit délinquant. Alors, il faut construire un modèle de prévision afin de déterminer les probabilités de délinquance de chacun des clients. Il existe plusieurs modèles pour effectuer une prévision et ceux-ci varient en complexité et en performance. Il s'agit donc de choisir celui qui convient le mieux à notre projet.

Les présentes analyses ont pour but de déterminer quelles méthodes ont un meilleur pouvoir prédictif sur la probabilité de délinquance d'un client. Nous allons commencer par faire les différentes analyses sous forme fréquentiste, (en se basant uniquement sur notre jeu de données des prêts investissements dans

la construction de notre modèle), et après avoir choisi la méthode optimale à l'aide de mesures statistiques, nous allons lui ajouter l'aspect bayésien. Dans ces analyses, nous avons tenté de prédire, avec le plus de justesse, la présence d'un retard futur chez les clients ayant un prêt investissement. Nous avons choisi le retard au lieu du défaut, car nous n'avons presque pas de défauts présents dans notre échantillon étudié, et par conséquent, il serait encore plus difficile de prédire celui-ci.

1.4.1. Régression logistique

Dans plusieurs champs d'étude, nous obtenons des variables dépendantes binaires, par exemple le succès ou l'échec, des variables dépendantes ordinales, par exemple faible, moyen, élevé et des variables dépendantes nominales, par exemple masculin ou féminin. Alors, la régression logistique est souvent utilisée pour modéliser une variable dépendante binaire avec un ensemble de variables explicatives, voir Hosmer et Lemeshow (2000).

Dans notre situation, nous avons une variable dépendante binaire, présence d'un retard ou non, que nous représentons par la variable Y. Ainsi, Y peut prendre uniquement deux valeurs, 0 et 1, 0 si le client n'est pas en retard et 1 s'il l'est, (les valeurs 0 ou 1 sont choisies arbitrairement). Si nous représentons l'ensemble de variables explicatives par le vecteur $\vec{x}$, nous voulons calculer la probabilité que Y soit égale à 1, ou $\pi = \Pr(Y = 1|\vec{x})$. En prédisant π, nous obtenons par le fait même, la probabilité complémentaire $1 - \pi = \Pr(Y = 0|\vec{x})$.

La fonction de vraisemblance à maximiser est donc :

$$\Pr(\vec{y}|\vec{x}, \vec{\beta}) = \prod_{i=1}^{n} \pi^{y_i} (1 - \pi)^{1-y_i},$$

avec le modèle logistique linéaire suivant :

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \vec{\beta}'\vec{x}, \tag{1.4.1}$$

où α est le paramètre constant et $\vec{\beta}$ est le vecteur des coefficients des variables explicatives. Le modèle, appelé « logit », cherche donc à optimiser la valeur de

ces deux paramètres en utilisant la méthode du maximum de vraisemblance, voir Rice (1995). De plus, nous pouvons aussi calculer les ratios pour chacune des variables explicatives. Ces ratios nous indiquent, pour le changement d'une unité de la variable explicative, la variation de la déliquance. Par exemple, pour un ratio égal à deux, cela signifie que pour l'augmentation d'une unité de cette variable, la déliquance double. Le ratio, lorsque la variable x_j passe de x_j à $x_j + 1$, se calcule de la façon suivante :

$$\text{ratio}_j = \exp(\beta_j), \qquad j \in 1, \ldots, M,$$

où la variable β_j provient du vecteur de l'équation (1.4.1) et M est le nombre de variables explicatives.

D'autres modèles peuvent être utilisées pour exprimer la probabilité π, comme la régression avec le modèle « probit ».

1.4.2. Régression avec modèle « probit »

Une régression avec le modèle « probit » ressemble beaucoup à la régression logistique de la section précédente, voir Finney (1971). La différence apparaît dans la prédiction de la variable Y ou dans le calcul de cette probabilité $\pi = \Pr(Y = 1|\vec{x})$. Le modèle général pour la régression possède la forme suivante :

$$\Pr(Y = 1|\vec{x}) = C + (1 - C)F(\vec{\beta}'\vec{x}),$$

où F est une fonction de distribution cumulative :

$$\Phi(x) \;=\; (2\pi)^{-0.5} \int_{-\infty}^{x} \exp(-z^2/2)dz,$$

et C est le taux de réponse naturel (la proportion des cas dans le groupe contrôle).

Ainsi, la fonction de vraisemblance à maximiser est :

$$\begin{aligned} L &= \prod_{y_i=0} \left[C + (1 - C)\Phi(\vec{\beta}'\vec{x}_i)\right] \prod_{y_i=1} \left[1 - C - (1 - C)\Phi(\vec{\beta}'\vec{x}_i)\right], \\ &= \prod_{y_i=0} \left[C + (1 - C)\Phi(\vec{\beta}'\vec{x}_i)\right] \prod_{y_i=1} \left[(1 - C)\left(1 - \Phi(\vec{\beta}'\vec{x}_i)\right)\right]. \end{aligned}$$

Cette fonction est maximisée en utilisant l'algorithme de Newton-Raphson, voir Bishop *et al.* (1975). Aussi, il est à noter qu'il existe d'autres fonctions de distribution cumulative, comme gompit/valeur extrême c'est-à-dire :

$$F(x) = 1 - \exp(-\exp(-x)).$$

Pour les besoins de ce mémoire, nous utilisons uniquement la fonction de distribution cumulative normale.

1.4.3. Analyse discriminante linéaire

Pour commencer, nous avons utilisé l'analyse discriminante linéaire afin de prédire la probabilité de retard des prêts investissements. Contrairement à la régression logistique qui modélise la probabilité π, l'analyse discriminante développe un critère discriminant directement des variables explicatives. À l'aide de ce critère, nous pouvons classer chacune des observations dans les groupes.

Nous avons choisi comme critère discriminant, une mesure de distance au carré généralisée, voir Rao (1973). Ainsi, l'analyse discriminante consiste à maximiser le ratio de la variance entre les groupes sur la variance à l'intérieur des groupes en utilisant une transformation linéaire. Alors, la fonction à maximiser est :

$$\begin{aligned} J(\varpi) &= \frac{\varpi^T S_B \varpi}{\varpi^T S_w \varpi}, \\ S_B &= \sum_k N_k (\mu_k - \bar{x})(\mu_k - \bar{x})^T, \\ S_W &= \sum_k \sum_{i \in k} (x_i - \mu_k)(x_i - \mu_k)^T, \end{aligned}$$

où S_B calcule la somme des différences entre les deux groupes et S_W calcule la somme des différences à l'intérieur des groupes et

$$\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{i \in k} x_i, \\ \bar{x} &= \frac{1}{N} \sum_i x_i = \frac{1}{N} \sum_k N_k \mu_k, \end{aligned}$$

N_k étant le nombre de cas dans la classe k. Alors, le but de l'analyse discriminante est de trouver une matrice de transformation ϖ qui définit la transformation

linéaire suivante :

$$\varpi^T : x \in R^{M\times 1} \rightarrow y = \varpi^T x \in R^{l\times 1}$$

où M est le nombre de variables explicatives et $l = C - 1$ où C est le nombre de classes. La matrice ϖ est composée des vecteurs propres de la solution du problème suivant :

$$S_W^{-1} S_B x = \lambda x.$$

Par contre, si le nombre de variables explicatives est trop grand comparé aux données, S_W est une matrice singulière et il faut utiliser un algorithme pour résoudre le problème des vecteurs propres, voir Howland, Jeon et Park (2003) pour cet algorithme. Alors, lorsque la multinormalité est supposée, une méthode paramétrique peut être développée comme fonction discriminante. Si nous ne pouvons pas supposer la multinormalité, des méthodes non paramétriques peuvent être utilisées. Ces méthodes comprennent, entre autres, le noyau et le $k^{\text{ième}}$ voisin plus proche, voir Rosenblatt (1956) et Parzen (1962).

1.4.4. Arbre de décision

Un arbre de décision est une représentation graphique d'un problème sous forme de branches et de noeuds. Ainsi, l'arbre de décision aide à séparer un problème complexe en plusieurs sous-ensembles avec des propriétés communes et, par le fait même, à en simplifier la prise de décision. En se basant sur de l'expérience et de l'information passées, il est donc possible de faire ressortir ces connaissances pour faciliter la prise de décision future.

L'arbre de décision se divise en deux catégories, soit :

- l'arbre de régression,
- l'arbre de classification.

L'arbre de régression modélise une variable dépendante réelle continue, par exemple, l'estimé du prix des maisons ou la longueur de temps d'une hospitalisation. L'arbre de classification modélise une variable dépendante catégorielle, par exemple le sexe (homme/femme) d'une personne ou le résultat d'une partie de baseball (perdu/gagné).

La méthode de classification par l'arbre de classification et de régression est représentée par l'acronyme « CART » et a été développée en 1984 par Breiman, Freidman, Olshen et Stone dans leur livre « *Classification and Regression Tree* » et résumée, plus tard, par Timofeev (2005).

Voici quelques caractéristiques de l'utilisation des arbres de classification :

Avantages :

1- C'est une méthode robuste et rapide pour les grosses bases de données.
2- Il est possible de valider le modèle à l'aide de tests statistiques.
3- C'est un modèle « boîte blanche », c'est-à-dire qu'il est simple et facile à interpréter.
4- Le modèle est capable de supporter les variables explicatives nominales et catégorielles, contrairement à d'autres méthodes qui n'utilisent qu'un des deux types.
5- Il n'est pas nécessaire de standardiser ou d'imputer les données manquantes.

Désavantage :

1- Il est facile d'obtenir un arbre avec beaucoup trop de noeuds, particulièrement dans les cas où nous n'avons pas beaucoup de données.

Après quelques études prélimiaires, nous avons constaté que l'arbre de classification, par ses avantages et ses mesures d'associations élevées, est la meilleure méthode parmi celles énoncées pour la modélisation du retard dans le remboursement d'un prêt investissement. Il est évident qu'il existe d'autres méthodes encore

plus performantes récemment développées, comme les réseaux de neurones, mais vu la simplicité de l'implantation des arbres de classification, nous avons décidé d'aller dans cette direction.

Il est à noter aussi que le choix de la meilleure méthode est corrélé avec les données, c'est-à-dire que la performance d'un modèle varie beaucoup selon la qualité et le type (qualitatif ou quantitatif) des variables explicatives. Ainsi, l'arbre de classification ne donne pas une classification systématiquement meilleure que la régression logistique ou l'analyse discriminante. Bref, vu que l'arbre de classification semble donner les meilleurs résultats pour les prêts investissements, nous allons développer cette méthode avec notre base de données. L'arbre de classification est développé en détails dans les chapitres 3 et 4.

Dans ce chapitre, nous avons présenté les deux objectifs de ce mémoire et nous avons donné les définitions financières et mathématiques utilisées dans les chapitres suivants. De plus, nous avons présenté trois tests statistiques, le test t, le test de la somme des rangs de Wilcoxon et le d de Cohen. Finalement, nous avons présenté trois modèles de prévision, soit la régression logistique, «probit», et l'analyse discriminante et nous avons introduit l'arbre de classification.

Chapitre 2

BASE DE DONNÉES

Dans un premier temps, nous avons récolté le plus d'informations possible sur nos clients afin de trouver les renseignements les plus utiles à notre prévision et à l'établissement d'un lien entre le fonds mutuel et le retard sur le remboursement d'un prêt investissement. Pour ce faire, nous avons regroupé ensemble les clients ayant un prêt investissement et leurs renseignements personnels récoltés à l'émission et au suivi du prêt dans une seule base de données. Ensuite, nous avons fait une extraction de la plus grande quantité d'informations *a priori* disponibles, relevées sur un autre type de prêt similaire aux prêts investissements et dont les clients ne sont pas nécessairement les mêmes, et nous avons regroupé aussi ces informations dans une seule base de données.

Les prochaines sections présentent la méthodologie suivie ainsi que les différentes analyses effectuées sur ces deux bases. La construction des bases de données a été faite avec la collaboration de l'équipe de la gestion du crédit aux particuliers de la Banque Nationale du Canada et avec l'aide du logiciel SAS version 9.1.

2.1. BASE DE DONNÉES *a priori*

La base de données *a priori* a été choisie en fonction de la disponibilité de l'information et de la proximité éventuelle de la population par rapport à notre population cible. Après discussion avec l'équipe de la Banque Nationale, nous en sommes venus à la conclusion que la population la plus près des prêts investissements est le prêt personnel avec une détention d'un fonds mutuel ou tout autre

type de placement hors fonds mutuels. Il est à noter que les placements des clients détenant un prêt personnel ne sont pas mis en garantie et que le prêt peut être, par exemple, pour une maison ou une voiture. De plus, pour cause de confidentialité, seulement les placements des clients de l'institution ont été analysés.

Ainsi, nous avons choisi de prélever un échantillon à six moments différents, aussi appelé fenêtre, afin d'analyser le comportement de la délinquance des clients et de leur fonds mutuel au cours des trois années disponibles. Nous avons choisi de couvrir un an par fenêtre. Donc, pour certaine période, les fenêtres se recoupent.

Voici la correspondance des dates à chacune des fenêtres :

fenêtre 1 : 1er novembre 2001 au 31 janvier 2002 ;
fenêtre 2 : 1er février 2002 au 31 janvier 2003 ;
fenêtre 3 : 1er mai 2002 au 30 avril 2003 ;
fenêtre 4 : 1er août 2002 au 31 juillet 2003 ;
fenêtre 5 : 1er novembre 2002 au 31 octobre 2003 ;
fenêtre 6 : 1er février 2003 au 31 janvier 2004.

Pour ces six fenêtres, nous avons obtenu principalement quatre informations :

- le nombre de jours de retard présents dans chacune des six fenêtres,
- le nombre de défauts présents à la fin de chacune des fenêtres,
- un score de risque à la date de la fin de la fenêtre,
- le montant investi en placement à la date de la fin de la fenêtre.

Afin de faciliter la comparaison entre plusieurs variables, nous avons choisi de standardiser le score de risque et le montant investi. De plus, nous avons relevé seulement le montant investi en placement par le client, car il est souvent très difficile, voire impossible, de saisir le risque global du portefeuille ainsi que leur rendement généré.

2.1.1. Statistiques descriptives

Nous avons calculé les statistiques descriptives des quatre variables mentionnées ci-haut que nous présentons comme suit pour cause de confidentialité. Ces quatres variables sont :

1- Retard : cette variable est qualitative dichotomique avec les valeurs 0 = non délinquant, 1 = délinquant.
2- Défaut : cette variable est qualitative dichotomique avec les valeurs 0 = non défaut, 1 = défaut.
3- Score de risque : cette variable est quantitative continue. De plus, cette variable possède une distribution unimodale avec asymétrie négative et elle conserve une distribution similiaire pour les 6 fenêtres.
4- Montant investi : cette variable est quantitative continue avec une distribution unimodale avec une asymétrie positive. Cette variable conserve une distribution similiaire pour les 6 fenêtres.

2.2. Base de données des prêts investissements

La base de données des prêts investissements est construite de façon à pouvoir obtenir un modèle afin de prédire le taux de retard des clients. Une première sélection de 34 variables sur les 220 a été conservée pour leur potentiel explicatif du retard et leur corrélation (voir annexe A pour un tableau résumé des corrélations).

Ainsi, l'objectif est de prédire la présence d'un retard (variable dichotomique) sur le remboursement d'un prêt investissement entre le 1er novembre 2005 et le 28 février 2006 à l'aide des 36 variables contenant les informations du 1er novembre 2004 au 31 octobre 2005. Habituellement, la prévision est effectuée sur le nombre de cas de défaut et non pas sur le nombre de retards. Dans notre situation, nous n'avons pas pu utiliser le défaut vu sa quasi non existence, alors nous avons décidé d'employer le retard. Il est à noter que toutes les méthodes énoncées sont applicables pour le défaut advenant un historique plus important.

TABLEAU. 2.1. Test t pour échantillons indépendants : Différence entre non retard et retard.

Variables	Valeur-p	Variables	Valeur-p
Age du client	0,142	Total engagement	0,623
Valeur prêt/valeur fonds	0,196	Retard 1 an	<0,001
Score risque actuel	<0,001	Retard temps	<0,001
Score risque précédent	<0,001	Retard fréquence	<0,001
Volatilité portefeuille	0,330	Rendement du fonds	0,965
Valeur marchande	0,372	novembre 2004	0,367
Valeur prêt-valeur fonds	0,836	décembre 2004	0,077
Score à l'octroi	<0,001	janvier 2005	0,207
Comportement octroi	0,893	février 2005	0,061
Variation prêt/fonds	0,355	mars 2005	0,097
Retard 90+jrs octroi	0,019	avril 2005	0,106
Retard -30jrs octroi	0,001	mai 2005	0,070
Retard -60jrs octroi	0,015	juin 2005	0,006
Retard -90jrs octroi	0,009	juillet 2005	0,010
Revenu principal	0,713	août 2005	0,037
Paiements mensuels (rotatif)	0,509	septembre 2005	<0,001
Écart-type Rendement	0,130	octobre 2005	<0,001

2.2.1. Statistiques descriptives

Dans le but de mieux visualiser et comprendre notre échantillon, nous avons fait quelques statistiques descriptives. Dans un premier temps, nous avons regardé, pour chacune des variables, si la moyenne de la variable est différente pour un client en retard d'un client non en retard. Pour ce faire, le test t pour échantillons indépendants est utilisé. Les différentes valeurs-p sont répertoriées dans le tableau 2.1.

Plusieurs variables sont sorties avec une valeur-p inférieure à 10 % de ce test. Nous avons choisi 10 % afin d'éviter l'élimination de variables possédant un pouvoir explicatif. Voici ces différentes variables :

(1) Score risque actuel, score risque précédent et score à l'octroi : ces variables sont quantitatives continues avec une distribution unimodale symétrique.

(2) Retard 90+jrs octroi, retard -30jrs octroi, retard -60jrs octroi, retard -90jrs octroi, retard 1 an, retard temps et retard fréquence : ces variables sont quantitatives continues avec une distribution unimodale et une asymétrie positive.

(3) Juin, juillet, août, septembre, octobre 2005 : ces variables sont quantitatives discrètes avec une distribution unimodal et une asymétrie positive. Ces variables représentent, historiquement, le nombre de jours de retards du client sur le remboursement de son prêt..

Après avoir considéré la multiplicité des tests, nous avons choisi 10 % comme seuil de signification. Ce seuil de signification est très conservateur, car nous voulons être certain de ne pas retirer une variable importante. En plus, nous conservons certaines variables qui n'ont pas une valeur-p inférieure à 10%, mais que nous pensons être très pertinentes :

(1) Âge du client : cette variable est quantitative continue et symétrique unimodale.

(2) Valeur prêt/valeur fonds : cette variable est quantitative continue avec une distribution unimodale et une asymétrie positive.

(3) Volatilité portefeuille : cette variable est quantitative continue avec une distribution unimodale et une asymétrie positive.

(4) Valeur marchande : cette variable est quantitative continue avec une distribution unimodale et une asymétrie positive.

(5) Valeur prêt-valeur fonds : cette variable est quantitative continue et symétrique unimodale.

(6) Comportement octroi : cette variable est quantitative continue avec une distribution bimodale.

(7) Variation prêt/fonds : cette variable est quantitative continue avec une distribution unimodale et une asymétrie positive.

(8) Revenu principal : cette variable est quantitative continue et symétrique unimodale.

(9) Paiements mensuels (rotatif) : cette variable est quantitative continue et symétrique unimodale.

(10) Écart-type rendement : cette variable est quantitative continue avec une distribution unimodale et une asymétrie négative.

(11) Total engagement : cette variable est quantitative continue avec une distribution unimodale et une asymétrie positive.

(12) Rendement du fonds : cette variable est quantitatve discrète avec une distribution unimodale et une asymétrie négative.

(13) Novembre, décembre 2004, janvier, février, mars, avril, mai 2005 : ces variables sont quantitatives discrètes avec une distribution unimodale et une asymétrie positive. Ces variables représentent, historiquement, le nombre de jours de retards du client sur le remboursement de son prêt.

Nous avons laissé tomber plusieurs variables parmi les 220, car leur coefficient de corrélation avec la variable de délinquance n'est pas assez important.

2.2.2. Standardisation et imputation

Nous avons choisi d'imputer La base de données des prêts investissements, car dans la plupart des analyses, le logiciel retire les observations contenant des données manquantes. Ainsi, ayant très peu de retards, nous risquons, sans imputation, de perdre des informations importantes. Un tableau de fréquences relatives et absolues des valeurs manquantes est présenté à l'annexe B. Il est à noter que

les variables n'ayant pas de valeurs manquantes sont omises dans ce tableau.

Avant d'imputer, nous avons décidé de comparer deux méthodes, l'imputation multiple et l'imputation par le plus proche voisin, afin de déterminer laquelle se rapproche le plus de la vraie valeur. Pour ce faire, nous prenons notre base de données sans les valeurs manquantes, et nous retirons, au hasard, plusieurs valeurs dans une des variables. Nous pouvons maintenant imputer ces fausses valeurs manquantes et les comparer avec les valeurs que nous avons retirées.

2.2.2.1. *L'imputation multiple*

Dans l'imputation multiple, chacune des valeurs manquantes est remplacée par un ensemble de $r(\geqslant 1)$ valeurs possibles pigées de la distribution prédictive *a posteriori*. La variation au travers des r imputations reflète l'incertitude créée par le remplacement des valeurs manquantes à partir des valeurs observées. Ainsi, après avoir appliqué l'imputation multiple, nous obtenons r jeux de données complets, où chacun peut être analysé. Lorsque les r analyses ont été effectuées, les résultats sont combinés avec la méthode énoncée par Rubin (1987). Dans notre situation, une seule valeur d'imputation ou $r = 1$ donne des résultats largement satisfaisants.

La distribution prédictive *a posteriori* des données manquantes ne se calcule pas analytiquement. Alors, pour résoudre ce problème, nous générons une chaîne de Monte Carlo markovienne ou « Markov Chain Monte Carlo (MCMC) » qui nous permet, après plusieurs itérations, d'approximer notre distribution prédictive *a posteriori*. En résumé, une chaîne de Markov est une séquence de variables aléatoires dans laquelle la distribution de chaque élément dépend de la valeur de l'élément précédent. Ainsi, en construisant une chaîne de Markov assez longue pour stabiliser l'échantillon à une distribution stationnaire, nous pouvons générer une série de nombres selon cette distribution en répétant cette chaîne plusieurs fois. Bref, un des avantages de l'imputation multiple est un maintien de la volatilité, car les valeurs imputées ne sont pas tirées des valeurs observées mais de leur

distribution.

L'imputation multiple a été introduite par Hastings (1970) et reprise par Li (1988) et Schafer (1997). Nous présentons, dans ce mémoire, un résumé de l'imputation mutiple en utilisant MCMC avec l'algorithme « Data Augmentation (DA) » de Tanner et Wong (1987).

Tout d'abord, pour simuler des observations indépendantes de notre distribution prédictive *a posteriori*, nous avons besoin d'introduire une distribution *a priori* ayant comme paramètre le vecteur ϑ. Alors, nous pouvons écrire l'équation suivante :

$$\Pr(x_{manq}|x_{obs}) = \int \Pr(x_{manq}|x_{obs}, \vartheta) \Pr(\vartheta|x_{obs}) d\vartheta, \tag{2.2.1}$$

où x_{manq} et x_{obs} correspondent respectivement à la partie manquante et observée de notre échantillon. Ensuite, nous devons supposer une distribution de probabilité à notre échantillon. Cette distribution est la plus souvent choisie comme étant la normale, Schafer(1997).

De plus, il faut connaître le lien qui existe entre les données manquantes et observées. Il existe trois types de liens :

(1) MAR : « Missing At Random », la probabilité de non-réponse est indépendante des données manquantes.

(2) MCAR : « Missing Completely At Random », la probabilité de non-réponse est indépendante des données observées et des données manquantes.

(3) NMAR : « Not Missing At Random », la probabilité de non-réponse est dépendante des données manquantes.

Ensuite, il faut générer des échantillons de l'équation (2.2.1). La probabilité $\Pr(\vartheta|x_{obs})$ est souvent difficile à simuler. Alors, nous « augmentons » la probabilité $\Pr(\vartheta|x_{obs})$ à $\Pr(\vartheta|x_{obs}, x_{manq})$ en supposant une valeur pour x_{manq}, d'où

l'utilisation de l'algorithme DA. Voici les étapes à suivre :

I- À l'aide d'une valeur initiale pour $\vartheta^{(t)}$, où t représente le numéro de l'itération, simuler une valeur pour la donnée manquante $x_{manq}^{(t+1)}$ à partir de la probabilité :

$$\Pr\left(x_{manq}|x_{obs},\vartheta^{(t)}\right).$$

P- À partir de cette valeur manquante $x_{manq}^{(t+1)}$, simuler une nouvelle valeur pour $\vartheta^{(t+1)}$ avec la probabilité :

$$\Pr\left(\vartheta|x_{obs},x_{manq}^{(t+1)}\right).$$

Nous appelons la première étape I parce qu'elle représente l'étape de l'Imputation et nous appellons la deuxième étape P parce qu'elle correspond à tirer ϑ de la distribution *a Posteriori* du jeu de données complet. En alternant ces deux étapes et en supposant que nos données sont MAR, la distribution des deux suites $\{\vartheta^{(t)};t=0,1,\ldots\}$ et $\{x_{manq}^{(t)};t=0,1,\ldots\}$ converge respectivement vers les probabilités $\Pr(\vartheta|x_{obs})$ et $\Pr(x_{manq}|x_{obs})$. Pour la valeur initiale $\vartheta^{(0)}$, Rubin (1987) propose de l'estimer avec l'algorithme « Expectation-Maximisation (EM) », car en plus de fournir une valeur initiale, cet algorithme nous donne une approximation du nombre d'itérations nécessaires. Il est prouvé dans Rubin (1987) que l'algorithme EM converge plus lentement que l'algorithme DA. Par conséquent, nous choisissons le nombre d'itérations de l'algorithme EM pour l'algorithme DA. Une description de l'algorithme EM peut être trouvée dans Dempster, Laird et Rubin (1977).

2.2.2.2. *Le plus proche voisin*

La méthode d'imputation par le plus proche voisin calcule la distance euclidienne entre chacune des observations. Ensuite, à l'aide des valeurs non manquantes des autres variables, elle impute la valeur manquante par l'observation ayant la distance euclidienne minimale. Un inconvénient de cette méthode est une perte de volatilité, car en imputant par une valeur qui est déjà présente dans le jeu de données, cela a pour effet de réduire sa volatilité. La méthode du plus

proche voisin est décrite, entre autres, par Spath (1980) ou Little et Rubin (1987).

Ainsi, nous remplaçons les valeurs manquantes par son voisin le plus proche en utilisant la distance euclidienne suivante :

$$d(m,i) = \sqrt{(x_1^m - x_1^i)^2 + \ldots + (x_{k-1}^m - x_{k-1}^i)^2 + (x_{k+1}^m - x_{k+1}^i)^2 + \ldots + (x_N^m - x_M^i)^2}, \tag{2.2.2}$$

où M est le nombre de variables explicatives, m représente l'observation avec une valeur manquante à la variable x_k, $i \in \{1, \ldots, m-1, m+1, \ldots, N\}$ indique l'observation et N est le nombre total d'observations. La variable x_k^i n'apparaît pas dans l'équation (2.2.2) puisque cette valeur est manquante. Alors, l'équation (2.2.2) calcule, pour la variable manquante, la distance entre les valeurs non manquantes des variables explicatives et les variables explicatives des autres observations :

$$d(m,1), \ldots, d(m,m-1), d(m,m+1), \ldots, d(m,N).$$

Par conséquent, si s est l'indice représentant la distance minimum, c'est-à-dire :

$$d(m,s) = \min_i d(m,i),$$

alors, nous imputons x_k^m par x_k^s. Si plusieurs distances sont égales, nous remplaçons la valeur manquante par la première distance trouvée ou le i le plus petit parmi les égalités.

2.2.2.3. *Comparaison*

Tout d'abord, nous construisons un histogramme afin de vérifier que les méthodes d'imputations respectent la distribution initiale. Nous observons à la figure 2.1 que les deux techniques suivent bien la distribution de la variable non imputée. Ensuite, nous vérifions la qualité de l'imputation en retirant au hasard des valeurs de la variable score du risque et en vérifiant, entre les différentes imputations, quelles méthodes s'en rapprochent le plus. En observant la figure 2.2, nous remarquons que l'imputation multiple fait mieux pour les données inférieures à environ 740 et que le plus proche voisin fait mieux pour celles supérieures.

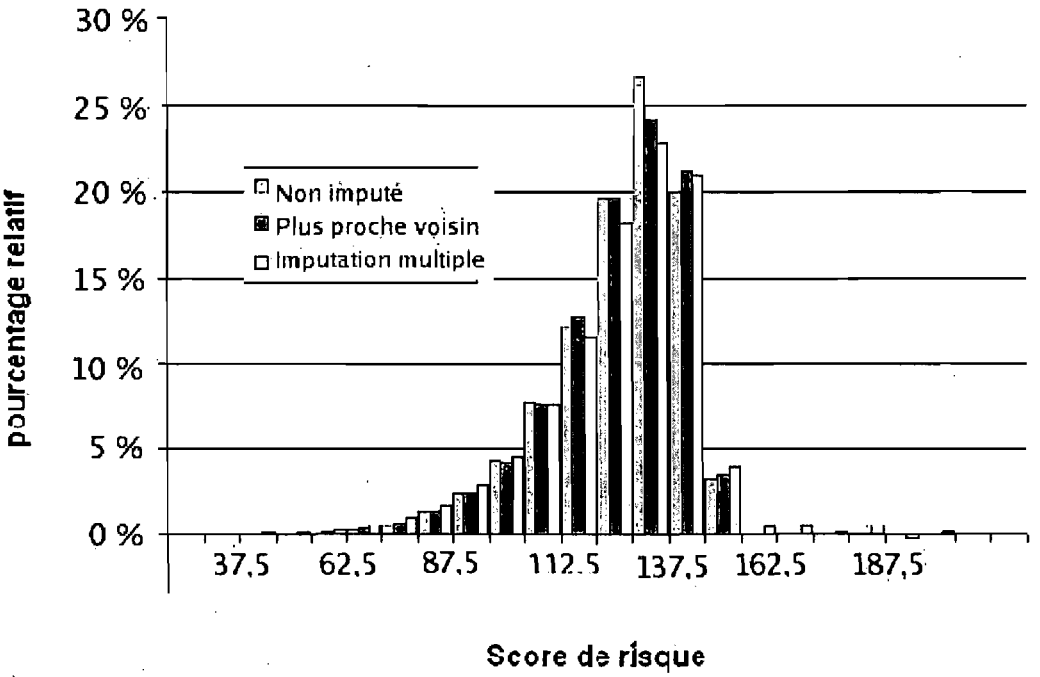


FIGURE. 2.1. Comparaison de la distribution imputée versus non imputée

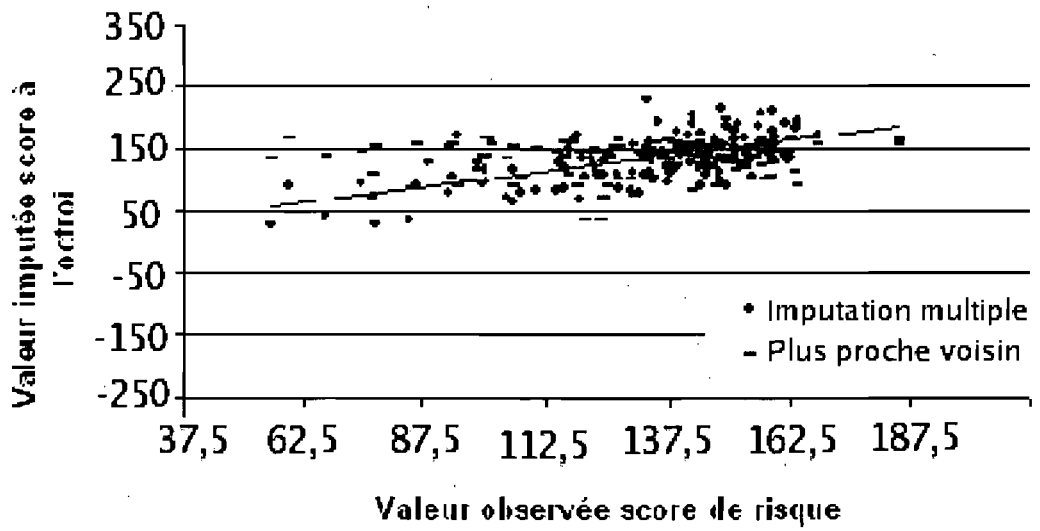


FIGURE. 2.2. Vérification de la qualité d'imputation

De plus, nous pouvons aussi calculer la distance avec la valeur réelle :

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N}(x_i^m - x_i^r)^2,$$

où x_i^r est la valeur réelle. Donc, plus $\bar{e}$ est proche de 0, plus l'imputation est bonne. Pour l'imputation multiple, nous avons $\bar{e} = 55{,}64$ ou en valeur relative 6,62 % et pour l'imputation par le plus proche voisin, nous avons $\bar{e} = 74{,}96$ ou en valeur relative 8,57 % . Dans la figure 2.2, la droite au milieu représente la valeur réelle du score de risques. Ainsi, la méthode choisie est l'imputation multiple car, en moyenne, elle est plus proche de la valeur réelle et elle cause une plus petite perte de volatilité des données.

2.3. Comportement des investisseurs

La situation qui nous intéresse est le comportement d'un placement lorsque le client devient délinquant sur son prêt. En d'autres mots, nous pouvons nous poser les deux questions suivantes :

1- Est-ce qu'un client qui possède un montant plus élevé en placements a moins de risques d'avoir un comportement délinquant qu'un client avec peu de placements ? Une autre avenue est d'analyser les capitaux propres (advenant la disponiblité de l'information).

2- Est-ce que les clients qui ont un rendement négatif (ou bien retirent simplement une partie de leur capital) sur leur placement ont tendance à être plus en retard que les clients avec un placement à capital croissant ?

Dans un premier temps, nous allons regarder si nous pouvons répondre par l'affirmative à ces deux questions pour les prêts personnels, dans un second temps, pour les prêts investisseurs et dans un dernier temps, advenant une conclusion

positive, nous allons utiliser cette information, dans les sections suivantes, pour notre modèle bayésien.

2.3.1. Question 1 : Est-ce qu'un client qui possède un montant plus élevé en placements a moins de risques d'avoir un comportement délinquant qu'un client avec peu de placements ?

Pour répondre à la question 1, nous avons regardé l'écart, en moyenne, entre la valeur des placements pour les clients en retard et les clients sans retard à chacune des fenêtres.

2.3.1.1. *Clients ayant un prêt personnel et un placement financier*

Une représentation graphique de la valeur marchande du fonds à la fin de la fenêtre en ordonnée (moyennes marginales) et de la date de la fenêtre en abcisse est donnée à la figure 2.3. Une moyenne marginale est la somme d'un des cas de la variable divisée par sa fréquence. La courbe du haut est celle des clients sans retard et celle du bas, des clients en retard. Nous avons alors vérifié si cet écart systématique entre les deux groupes est significatif. Vu la présence d'interaction entre le groupe et la date de la fenêtre, nous avons fait un test t pour échantillons indépendants à chacune des 6 fenêtres. En regardant le tableau 2.2, nous pouvons constater qu'il existe une différence significative entre les deux groupes à chacun des temps. Il est à noter ici qu'une correction pour la multiciplicité des tests n'est pas nécessaire vu que la valeur-p est inférieure à 0,001.

En regardant de plus près le tableau 2.2, il est clair que la relation est importante, car avec un degré de liberté aussi élevé, la plus petite différence devient significative. Étant donné que ce test dépend de la taille échantillonale, nous devons retirer cet effet de notre test. Alors, à l'aide de la mesure du d de Cohen présenté dans la section des préliminaires, nous pouvons déterminer la force de la relation sans l'effet taille. Les résultats sont présentés dans le tableau 2.2. Ainsi, en se référant au tableau 1.1, la différence entre les deux groupes de clients au niveau de la valeur marchande du fonds est déclarée comme moyenne.

De plus, nous pouvons faire une régression logistique avec comme variable dépendante la délinquance et comme variable explicative la valeur marchande du fonds. En catégorisant notre variable « montant investi », le calcul du ratio nous donne, par rapport à la catégorie de référence, la proportion de la délinquance. Par exemple, les clients avec des fonds inférieurs à 1 000 pour la fenêtre 1 ont 2,078 fois plus de chance d'être en retard que les clients avec des fonds au dessus de 150 000. Les ratios pour les prêts personnels sont présentés au tableau 2.3 et les chiffres en gras sont les ratios significatifs au niveau 0,1% (corrigé pour la multiplicité des tests).

Bref, pour répondre à la question 1, en moyenne, nous observons que les clients ayant de petits fonds ont environ 2 fois plus de chance d'être en retard que les clients avec des fonds de plus de 150 000.

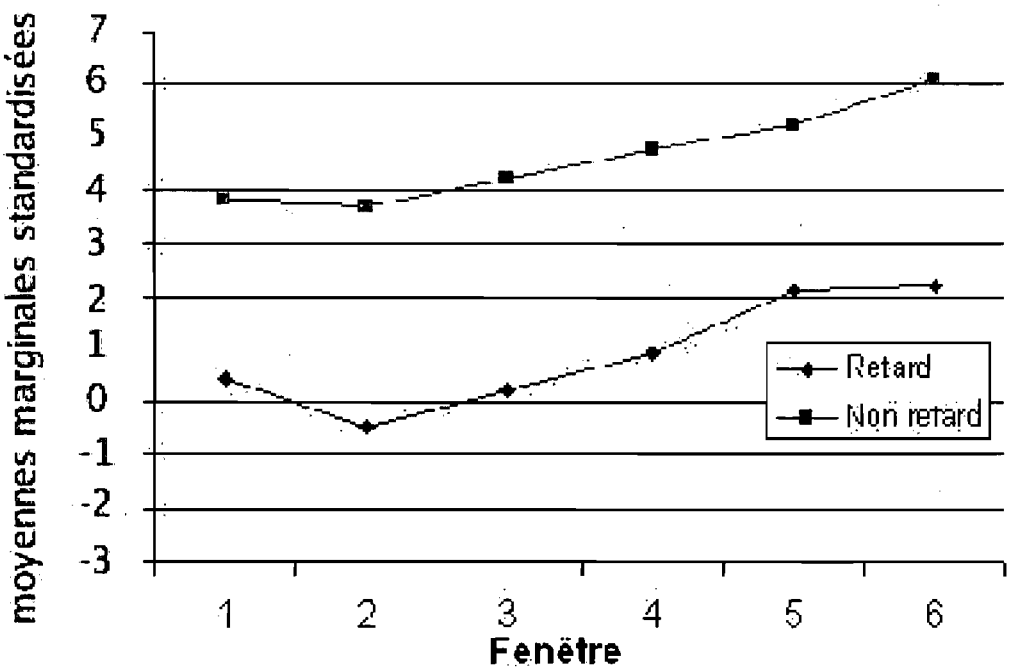


FIGURE. 2.3. Comportement du fonds à travers le temps

TABLEAU. 2.2. Test t pour échantillons indépendants et d Cohen pour la valeur du fonds

Fenêtre	Statistique t	degré de liberté	valeur-p	d Cohen
Fenêtre 1	3,0514	1724	0,002	0,456
Fenêtre 2	6,361	3828	<0,001	0,519
Fenêtre 3	5,679	3283	<0,001	0,487
Fenêtre 4	5,3144	3219	<0,001	0,448
Fenêtre 5	4,153	3020	<0,001	0,346
Fenêtre 6	5,178	3006	<0,001	0,405

TABLEAU. 2.3. Ratio aux différentes fenêtres pour le montant investi

Valeur du fonds	Fen. 1	Fen. 2	Fen. 3	Fen. 4	Fen. 5	Fen. 6
< 1 000	**2,078**	**2,500**	**2,011**	**1,751**	**1,953**	**2,082**
1 000 - 5 000	1,356	1,607	1,114	1,121	1,076	1,469
5 000 - 10 000	1,015	1,412	1,022	0,980	0,953	1,299
10 000 - 20 000	0,942	1,125	0,840	0,833	0,805	1,007
20 000 - 30 000	0,654	0,933	0,812	0,791	0,760	1,041
30 000 - 40 000	0,831	0,919	0,583	0,549	**0,472**	0,698
40 000 - 50 000	0,718	0,887	0,543	0,453	**0,444**	0,691
50 000 - 60 000	0,631	0,799	0,575	0,554	0,549	0,726
60 000 - 70 000	0,822	0,471	0,544	0,429	**0,438**	0,581
70 000 - 100 000	0,847	0,972	0,637	0,753	0,836	1,126
100 000 - 130 000	0,691	0,917	0,677	0,733	0,896	1,243
130 000 - 150 000	0,842	1,528	0,963	0,900	0,739	1,049
>150 000	catégorie de référence					

2.3.1.2. *Clients avec un prêt investissement*

Pour les prêts investissements, nous avons uniquement la valeur marchande du fonds à une seule fenêtre, soit 28 février 2006, donc nous ne pouvons faire de graphique comme dans le cas précédent (figure 2.3). Par contre, nous avons fait un test t pour échantillons indépendants, entre la valeur marchande des clients

sans retard et la valeur marchande des clients en retard (valeur-p=0,372, voir tableau 2.1). La valeur-p étant supérieure à 0,05, nous pouvons conclure que, pour les prêts investissements, il ne semblerait pas y avoir de différences significatives en moyenne entre la valeur marchande du fonds des clients en retard sur le remboursement de leur prêt et ceux sans retard.

Bref, nous émettons l'hypothèse qu'il devrait exister une différence en moyenne de la valeur marchande du fonds entre les clients en retard et sans retard même si celle-ci ne semble pas significative. Nous justifions cette hypothèse par le fait que les prêts personnels possèdent cette relation et ceux-ci semblent s'apparenter aux prêts investissements. Cette hypothèse pourra certainement être confirmée dans le futur avec l'ajout de nouveaux cas de retards.

2.3.2. Question 2 : Est-ce que les clients qui ont un rendement négatif (ou bien retirent simplement une partie de leur capital) sur leur placement ont tendance à être plus en retard que les clients avec un placement à capital croissant ?

Pour répondre à la question 2, nous avons analysé le rendement du placement plutôt que le montant lui-même, c'est-à-dire :

$$r_t = \frac{v_t - v_{t-1}}{v_t},$$

où r_t est le rendement au temps t, v_t est la valeur du fonds au temps t et $t \in \{1, \ldots, 6\}$ est la fenêtre. Ainsi, nous avons regardé l'écart, en moyenne, entre le rendement des placements pour les clients en retard et les clients sans retard à chacune des fenêtres.

2.3.2.1. *Clients ayant un prêt personnel et un placement financier*

Nous avons représenté le rendement du fonds à la fin de la fenêtre en ordonnée et de la date de la fenêtre en abcisse à la figure 2.4. Ensuite, nous avons vérifié si cet écart entre les deux groupes est significatif. Comme pour la valeur du fonds, nous avons regardé le test t pour échantillions indépendants et la mesure du d de Cohen. Les résultats pour le rendement sont présentés au tableau 2.4. Donc, il

TABLEAU. 2.4. Test t pour échantillons indépendants et d Cohen pour le rendement

Fenêtre	Statistique t	degré de liberté	valeur-p	d Cohen
Fenêtre 1-2	1,002	26745	0,317	0,071
Fenêtre 2-3	-1,298	33106	0,194	0,086
Fenêtre 3-4	0,542	35946	0,588	0,039
Fenêtre 4-5	0,388	34798	0,698	0,017
Fenêtre 5-6	0,741	33908	0,459	0,085

ne semble pas y avoir de lien entre le rendement et le retard au niveau des prêts personnels avec un placement financier.

La courbe du rendement ne suit pas celle de la valeur du fonds puisque nous avons un nombre restreint de prêts dont le rendement est disponible. Cette perte en terme de nombre de prêts pour le rendement provient du fait que le prêt doit être ouvert pendant toute la période couverte par la fenêtre, ce qui n'est pas le cas pour la valeur du fonds.

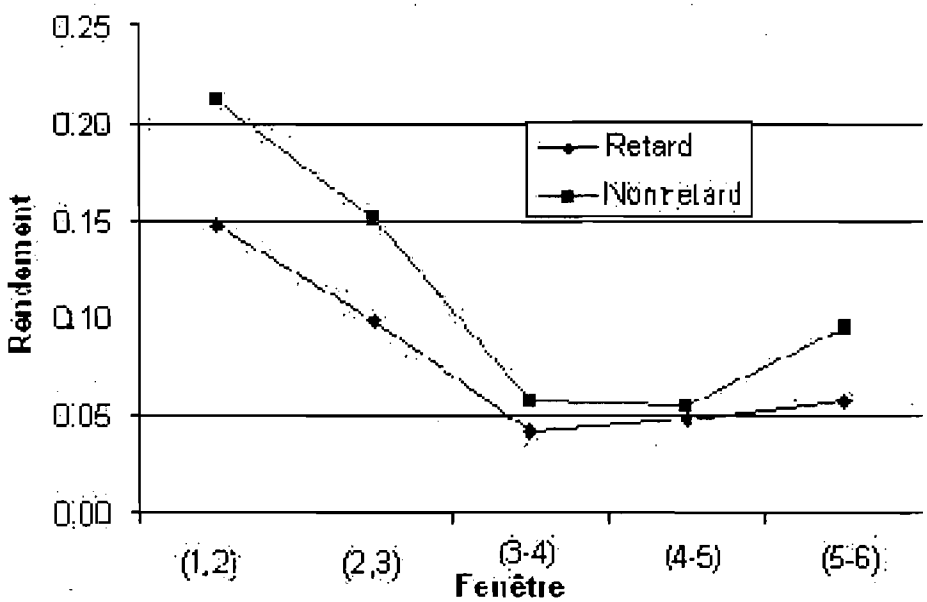


FIGURE. 2.4. Comportement du rendement à travers le temps

2.3.2.2. *Clients avec un prêt investissement*

Pour les prêts investissements, nous avons uniquement le rendement du fonds à une seule fenêtre, soit 28 février 2006, donc nous ne pouvons faire de graphique comme dans le cas précédent. Par contre, comme pour la question 1, nous avons fait un test t pour échantillons indépendants, entre le rendement des fonds des clients sans retard et le rendement des fonds des clients en retard (valeur-p=0,965, voir tableau 2.1). La valeur-p étant supérieure à 0,05, nous pouvons conclure que, pour les prêts investissements, il ne semblerait pas y avoir de relation entre le rendement du fonds et le retard.

Bref, le rendement n'est pas significatif pour les prêts personnels et les prêts investissements. Il serait donc injustifié d'utiliser cette information comme *a priori*.

Dans ce chapitre, nous avons montré les statistiques descriptives des deux bases de données et nous avons trouvé une information *a priori* pour les prêts investissements, soit le lien entre la valeur du fonds et la délinquance. Ce lien est implanté dans le modèle par la méthode du chapitre 4 et les résultats sont présentés dans le chapitre 5. Nous poursuivons, dans le chapitre suivant, avec la méthodologie détaillée sur les arbres de classification.

Chapitre 3

ARBRE DE CLASSIFICATION FRÉQUENTISTE

Dans ce mémoire, nous avons choisi d'expliciter l'arbre de classification en détail puisque notre variable dépendante est une variable dichotomique. Il est à noter qu'une méthode similaire peut être obtenue pour une variable dépendante continue, soit l'arbre de régression. Nous référons le lecteur à Breiman *et al.* (1984) pour l'algorithme complet sur l'arbre de régression. Dans ce chapitre, nous illustrons la procédure de la construction d'un arbre de classification basée sur la thèse de Timofeev (2005).

Dans les prochaines sections, nous abordons la préparation de la base, l'objectif de l'arbre de classification ainsi que les définitions nécessaires. Nous énonçons les régles de séparation utilisées pour les noeuds. Nous construisons l'algorithme complet permettant l'achèvement de l'arbre avec ses différentes étapes, comme le choix du nombre de noeuds terminaux, l'élagage de l'arbre (opération d'arboriculture qui consiste à retrancher d'un arbre les branches superflues et nuisibles) et finalement, la procédure à suivre lors de la présence de valeurs manquantes.

3.1. PRÉPARATION DE LA BASE DE DONNÉES ET DÉFINITIONS

Tout d'abord, l'échantillon à modéliser est séparé en deux, c'est-à-dire l'échantillon d'apprentissage et l'échantillon d'évaluation:

Il faut faire attention lors de la séparation, car l'échantillon d'évaluation doit être considéré comme indépendant de l'échantillon d'apprentissage. La procédure la plus commune est de tirer au hasard l'échantillon d'évaluation de notre base de données. Historiquement, l'échantillon d'évaluation représente un tiers de l'échantillon, mais il n'existe pas de justification théorique à cette séparation. L'échantillon d'évaluation nous sert dans le calcul de la « vraie puissance prédictive » de l'arbre. L'arbre de classification n'est utilisé seulement lorsque, pour chaque observation de l'échantillon, nous connaissons la classe à laquelle l'observation appartient (appelé apprentissage supervisé). Dans notre étude, la variable dépendante possède uniquement deux classes. La classe 1 est définie comme étant les clients sans retard sur le remboursement de leur prêt et la classe 2 comme étant les clients avec un ou plusieurs retards.

Ensuite, après avoir séparé notre base en l'échantillon d'apprentissage et d'évaluation, voici les deux étapes à suivre :

1- construction d'un arbre sur l'échantillon d'apprentissage,
2- évaluation de la qualité de la prévision de l'arbre avec l'échantillon d'évaluation.

Ainsi, nous allons conserver l'arbre qui possède la meilleure prévision sur notre échantillon d'évaluation.

Objectif d'un arbre de classification

L'objectif d'un arbre de classification est de diviser l'échantillon en plusieurs sous-échantillons homogènes. Alors, à l'aide de séparations optimales, l'arbre évolue en séparant la base de données en branches de plus en plus fines jusqu'à ce que les noeuds soient homogènes ou jusqu'à ce que la règle d'arrêt soit satisfaite (voir la section 3.4 pour plus de détails sur la règle d'arrêt).

Définition 3.1.1 (Le noeud). *Le noeud d'un arbre représente graphiquement un ensemble de données. À chaque fois que cet ensemble de données (ou ce noeud) est subdivisé en deux par une règle de séparation, il y a création de deux autres noeuds représentant respectivement ces nouveaux ensembles de données.*

Définition 3.1.2 (Une branche). *Une branche d'un arbre de classification est composée d'un noeud d'origine et de l'ensemble des noeuds découlant de ce noeud d'origine.*

Définition 3.1.3 (Noeud terminal). *Un noeud est dit terminal lorsque la règle d'arrêt est satisfaite ou lorsqu'il n'est plus possible de le séparer (dans le cas d'un noeud homogène).*

Définition 3.1.4 (La profondeur d'un noeud). *La profondeur d'un noeud se définit comme étant le nombre de conditions à satisfaire pour se retrouver dans ce noeud.*

Définition 3.1.5 (L'homogénéité parfaite). *L'homogénéité parfaite d'un noeud est définie comme étant la présence, dans ce noeud, d'une seule classe (une classe est composée d'éléments identiques uniquement ou d'un intervalle disjoint). Par exemple, si nous cherchons à modéliser la présence d'une maladie versus son absence, un noeud homogène contiendrait uniquement des patients atteints ou uniquement des patients non atteints.*

Il s'agit maintenant de traduire ces définitions en langage mathématique et de le résoudre par un algorithme détaillé. Alors, voici les différentes notations que nous allons employer lors des prochaines sections :

- t_p = noeud parent,
- t_l, t_r = noeud enfant gauche et droite respectivement,
- P_l, P_r = probabilité des noeuds de gauche et de droite,

- $p(k|t)$ = probabilité d'être dans la classe k sachant que nous sommes dans le noeud t,
- x_j = variable dépendante j,
- x_j^R = meilleure valeur de séparation pour x_j.

À l'aide de ces définitions, l'arbre de classification peut être maintenant représenté visuellement par le diagramme orienté de la figure 3.1. Le noeud parent est

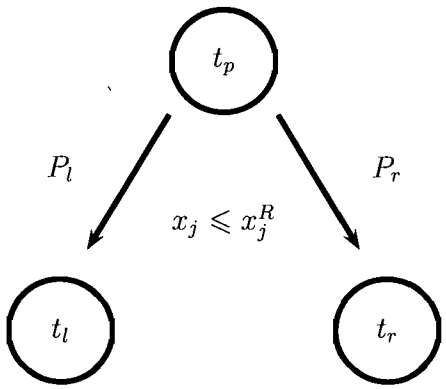


FIGURE. 3.1. Arbre de classification

séparé par une règle $x_j \leqslant x_j^R$ avec probabilité P_R d'aller à droite et P_L d'aller à gauche d'où la naissance de deux autres noeuds t_l et t_r. Ainsi, la construction d'un arbre de classification consiste en 4 éléments importants :

- un ensemble de questions Q, de la forme $x_j \leqslant x_j^R$,
- un critère de séparation pouvant être utilisé pour chacun des noeuds et des séparations,
- la décision de déclarer un noeud terminal ou de continuer à le séparer,
- la désignation d'une classe à chacun des noeuds.

De plus, l'ensemble des questions doivent respecter ces trois points :

- chaque séparation dépend de la valeur d'une seule variable,

- pour chaque variable ordonnée, l'ensemble Q comprend les questions de la forme $x_j \leqslant c, c \in (-\infty, \infty)$,
- si x_j est nominal, prenant les valeurs, par exemple $\{b_1, b_2, \ldots, b_\kappa\}$ où κ est le nombre de valeurs différentes, alors l'ensemble Q comprend toutes les questions de la forme $x_j \in S$ où S est un sous-ensemble de $\{b_1, b_2, \ldots, b_L\}$.

Par la suite, pour chaque noeud, l'algorithme cherche au travers de chacune des variables la meilleure séparation, en commençant par la variable x_1 jusqu'à x_M où M est le nombre total de variables. Pour connaître la meilleure séparation, nous définissons l'homogénéité d'un noeud en particulier comme étant la fonction d'impureté $i(t)$, où t est le noeud. L'impureté est nulle lorsque le noeud atteint l'homogénéité parfaite. Nous pouvons définir l'impureté totale comme étant la somme de l'impureté à chacun des noeuds terminaux $\sum_{t \in \text{ noeuds terminaux}} i(t)$. Par conséquent, l'objectif est de minimiser l'impureté totale de l'arbre. Comme la fonction d'impureté est constante pour les noeuds parents puisque la règle de séparation affecte seulement les noeuds enfants et ce, peu importe la séparation $x_j \leqslant x_j^R$, l'homogénéité maximale des noeuds enfants gauche et droite est équivalente au changement de la fonction d'impureté, notée $\Delta i(s,t)$:

$$\Delta i(s,t) = i(t_p) - E[i(t_c)], \tag{3.1.1}$$

où t_c sont les noeuds enfants gauche et droite, t_p est le noeud parent et $s : x_j \leqslant x_j^R$ est la règle de séparation. En assumant que P_l, P_r sont les probabilités des noeuds de gauche et de droite, nous pouvons réécrire l'équation (3.1.1) sous la forme :

$$\Delta i(s,t) = i(t_p) - P_l i(t_l) - P_r i(t_r). \tag{3.1.2}$$

Vu que $i(t_p)$, l'impureté du noeud parent, demeure fixe, cela revient à minimiser l'impureté des noeuds enfants. Donc, pour chaque noeud, il faut résoudre le problème de maximisation suivant :

$$\underset{x_j \leqslant x_j^R, j=1,\ldots M,}{\arg\max} \left[i(t_p) - P_l i(t_l) - P_r i(t_r) \right]. \tag{3.1.3}$$

L'équation (3.1.3) signifie que la méthode va passer au travers de toutes les valeurs possibles et de toutes les variables pour trouver la règle $x_j \leqslant x_j^R$ qui

va maximiser le changement d'impureté $\Delta i(t)$. Maintenant, il faut se demander comment allons-nous définir la fonction d'impureté $i(t)$? En théorie, il existe plusieurs fonctions d'impureté, mais dans ce mémoire, nous allons présenter les deux plus utilisées en pratique, soit la règle de séparation de Gini et la règle de séparation de Twoing, voir Timofeev (2005).

3.2. Règle de séparation

3.2.1. Règle de séparation de Gini

La règle de séparation de Gini pour un noeud en particulier consiste en la multiplication de la probabilité d'appartenir à une classe pour chacune des paires de classes possibles. Ainsi, pour une homogénéité parfaite, la probabilité d'être dans cette classe est 1 et dans les autres, nulle. Alors, la fonction d'impureté pour un noeud à deux classes dans une homogénéité parfaite est $1 \times 0 + 0 \times 1 = 0$ (paires possibles (1,2) et (2,1)). Dans l'autre extrême, où toutes les classes auraient une probabilité égale, nous obtenons une impureté maximale, comme pour un noeud à deux classes, $0{,}5 \times 0{,}5 + 0{,}5 \times 0{,}5 = 0{,}5$. Ainsi dans cet exemple, la fonction d'impureté est représentée à la figure 3.2. Donc, de façon générale pour plusieurs

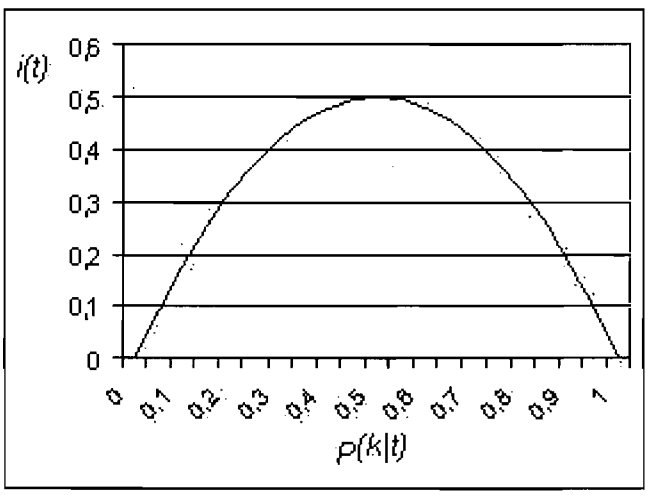


FIGURE. 3.2. Graphique de la fonction d'impureté pour un noeud à deux classes

classes, la règle de séparation de Gini utilise la fonction d'impureté $i(t)$ suivante :

$$\begin{aligned} i(t) &= \sum_{k \neq l} p(k|t)\, p(l|t) \\ &= \sum_{k} p(k|t) \sum_{l \neq k} p(l|t) \\ &= \sum_{k} p(k|t)\,(1 - p(k|t)) \\ &= 1 - \sum_{k} p^2(k|t), \end{aligned} \tag{3.2.1}$$

où $k, l = 1, \ldots C$ sont les indices de classes et $p(k|t)$ est la probabilité conditionnelle que la classe k provienne du noeud t. Dans notre cas, la variable C prend la valeur 2 puisque nous avons seulement deux classes, sans retard et avec retard. Nous conservons C dans les équations subséquentes dans le but d'obtenir une forme générale de la règle de séparation. Ainsi, en appliquant l'équation (3.2.1) à (3.1.2), nous obtenons la fonction de changement d'impureté $i(t)$ suivante :

$$\Delta i(s,t) = -\sum_{k=1}^{C} p^2(k|t_p) + P_l \sum_{k=1}^{C} p^2(k|t_l) + P_r \sum_{k=1}^{C} p^2(k|t_r). \tag{3.2.2}$$

Les probabilités P_l et P_r sont complémentaires, c'est-à-dire $P_l + P_r = 1$. En utilisant cette relation, nous réécrivons l'équation (3.2.2) de cette façon :

$$\Delta i(s,t) = P_l \sum_{k=1}^{C} \left[p^2(k|t_l) - p^2(k|t_p)\right] + P_r \sum_{k=1}^{C} \left[p^2(k|t_r) - p^2(k|t_p)\right],$$

et le problème d'optimisation devient alors,

$$\underset{x_j \leqslant x_j^R, j=1,\ldots M,}{\arg\max} \left[P_l \sum_{k=1}^{C} \left[p^2(k|t_l) - p^2(k|t_p)\right] + P_r \sum_{k=1}^{C} \left[p^2(k|t_r) - p^2(k|t_p)\right] \right].$$

La règle de séparation de Gini cherche, au travers de l'échantillon d'apprentissage, la classe ayant le plus faible degré d'impureté en utilisant les probabilités conditionnelles de celle-ci. Cette méthode fonctionne bien pour des données avec du bruit.

3.2.2. Règle de séparation Twoing

Contrairement à la règle de séparation de Gini qui consiste à minimiser le produit de ces deux probabilités $p(1|t), p(2|t)$, la règle de séparation de Twoing

cherche à minimiser le produit de ces deux probabilités $p(1|t_l), p(1|t_r)$. Ainsi, la règle de séparation de Twoing conserve la même définition de l'impureté, mais l'équation de changement d'impureté devient la suivante :

$$\Delta i(s,t) = \frac{P_l P_r}{4}\left[\sum_{k=1}^{C} |p(k|t_l) - p(k|t_r)|\right]^2,$$

ce qui implique le nouveau problème d'optimisation suivant :

$$\underset{x_j \leqslant x_j^R, j=1,\ldots M,}{\arg\max} \frac{P_l P_r}{4}\left[\sum_{k=1}^{C} |p(k|t_l) - p(k|t_r)|\right]^2.$$

L'inconvénient des règles de séparation de Twoing et de Gini est que ces procédures sont longues à optimiser, car le nombre de séparations possibles augmente rapidement avec la quantité d'information disponible. La règle de séparation de Twoing est idéale lorsque nous avons un grand nombre de classes à prédire.

3.2.3. Autres règles de séparation

Il existe plusieurs autres méthodes de séparation :

(1) la règle de séparation par l'entropie,

(2) la règle du khi-deux,

(3) la règle de la déviation maximale,

mais il a été prouvé par Breiman *et al.* (1984) que l'arbre final ne dépend pas réellement du choix de la règle de séparation. Lors de la construction d'un arbre de classification, c'est la procédure de réduction ou d'élagage qui est la plus importante.

3.3. ALGORITHME

Il faut, tout d'abord, définir les probabilités pour chacune des classes du noeud parent. Ensuite, à la première valeur de la variable à séparer, il faut définir la probabilité d'aller à gauche, à droite et la probabilité d'obtenir absent/présent dans ces noeuds. Dans ce mémoire, nous nous intéressons seulement au cas des deux classes et laissons le lecteur généraliser la méthode pour plusieurs classes.

Ainsi, vu que nous n'avons que deux classes, nous pouvons réécrire la fonction d'impureté de Gini par :

$$i(t) = p(1|t)\,p(2|t) + p(2|t)\,p(1|t) = 2p(1|t)\,p(2|t).$$

Ainsi, le changement d'impureté devient :

$$\begin{aligned}\Delta i(s,t) &= i(t_p) - P_l i(t_l) - P_r i(t_r), \\ &= 2p(1|t_p)\,p(2|t_p) - 2P_l p(1|t_l)\,p(1|t_r) - 2P_r p(1|t_r)\,p(2|t_r).\end{aligned}$$

La fonction d'impureté de Twoing n'est pas reliée à l'impureté d'un noeud, mais directement au changement d'impureté. Dans le cas de deux classes, nous pouvons définir le changement d'impureté par :

$$\Delta i(s,t) = \frac{P_l P_r}{4}\left[|p(1|t_l) - p(1|t_r)| + |p(2|t_l) - p(2|t_r)|\right]^2.$$

Voici les grandes étapes de la méthode à suivre pour construire un arbre de classification.

(1) Ordonner l'échantillon et prendre le minimum de la première variable comme valeur de séparation.

(2) Calculer, pour chacune des classes du noeud parent, leur fréquence.

(3) Calculer, pour chacune des classes du noeud enfant gauche/droite, leur fréquence.

(4) Calculer le changement d'impureté.

(5) Recommencer les étapes 1 à 4 en changeant à l'étape 1 la valeur de séparation par la seconde valeur la plus près du minimum et ainsi de suite pour toutes les valeurs différentes de la première variable.

(6) Trouver le maximum du changement d'impureté de cette variable.

(7) Répéter 1 à 6 pour chacune des variables explicatives.

(8) Trouver le maximum de changement d'impureté parmi toutes les variables et conserver cette séparation comme premier noeud parent.

(9) Répéter 1 à 8, en partant de chacun des noeuds enfants générés par la séparation, jusqu'à ce que le changement d'impureté soit nul ou que celui-ci soit égal entre les différentes variables.

L'étape (9) s'appelle la règle d'arrêt. Le calcul de la règle de séparation de Gini est simple et facilement programmable, tout comme la règle de séparation de Twoing d'où leur fréquente utilisation.

3.4. Choix du nombre de noeuds terminaux

La règle d'arrêt, énoncée dans le point précédent, peut constituer un problème lorsque nous avons affaire à un jeu de données plus ou moins important. La complexité de l'arbre peut atteindre des centaines de niveaux et contenir des noeuds non significatifs. Il faut donc appliquer un algorithme d'élagage ou de réduction de taille. Il existe, entre autres, trois méthodes permettant de réduire l'arbre de classification, le changement minimal d'impureté, l'optimisation par le nombre minimal de points et le meilleur arbre de Breiman.

3.4.1. Changement maximal d'impureté

Une méthode fort simple est d'imposer un changement minimal d'impureté, c'est-à-dire que si :

$$\max_{x_j \leqslant x_j^R,\ j=1,\ldots M,} \Delta i\,(s,t) < \beta, \qquad \text{où } \beta > 0$$

alors le noeud est terminal. Plusieurs problèmes surviennent en utilisant cette méthode :

(1) le choix de β est arbitraire,

(2) il peut exister un noeud tel que $\max_{x_j \leqslant x_j^R,\ j=1,\ldots M,} \Delta i\,(s,t)$ est petit, mais que les noeuds descendants de celui-ci peuvent avoir de grandes décroissances d'impuretés,

(3) si β est trop bas, l'arbre est trop complexe.

Une autre méthode consiste en l'optimisation par le nombre minimal de points.

3.4.2. L'optimisation par le nombre minimal de points

Cette méthode possède, comme la précédente, l'avantage d'être simple. Elle consiste à arrêter la séparation lorsque le nombre d'observations présentes dans un noeud est inférieur à un certain pourcentage. Cette approche marche rapidement et elle s'applique facilement. Par contre, comme la méthode précédente, le choix du pourcentage est arbitraire. Dans la littérature, le pourcentage utilisé le plus fréquemment est 10 % de l'échantillon d'apprentissage. Les méthodes du changement minimal d'impuretés et de l'optimisation par le nombre minimal de points sont fort simples, mais elles génèrent souvent des arbres trop ou insuffisamment complexes.

3.4.3. Le meilleur arbre (Breiman *et al.*, 1984)

Nous pouvons remarquer que si nous diminuons le pourcentage minimal dans un noeud terminal ou le β du changement d'impureté, nous augmentons la complexité de l'arbre et nous diminuons son impureté. Par contre, en augmentant le pourcentage minimal et le β, nous diminuons sa complexité et augmentons son impureté. La question est : quel est le meilleur compromis entre la complexité et l'impureté ?

Cette méthode est basée sur la proportion optimale entre la complexité et l'erreur de mauvaise classification. La complexité de l'arbre de décision donne de mauvais résultats sur des échantillons indépendants. La performance sur ces échantillons est nommée la « vraie puissance prédictive » de l'arbre.

Ainsi, le rapport optimal entre la complexité et l'erreur de mauvaise classification est trouvé par la fonction suivante :

$$\arg\min_{T} R_O(T) = \arg\min_{T} \left(R(T) + O\left(\tilde{T}\right) \right), \tag{3.4.1}$$

où $R(T)$ est l'erreur de mauvaise classification de l'arbre T et $O\left(\tilde{T}\right)$ est la mesure de complexité dépendant de $\tilde{T}$ (l'ensemble des noeuds terminaux de l'arbre).

Ainsi, la fonction R peut s'écrire sous cette forme :

$$\begin{aligned} R(T) &= \sum_{t\in\tilde{T}} r(t)\, p(t) = \sum_{t\in\tilde{T}} (1 - \max[p(1|t), p(2|t)]) \times (N(t)/N), \\ R(\{t\}) &= r(t)\, p(t), \\ r(t) &= 1 - \max[p(1|t), p(2|t)], \\ p(t) &= N(t)/N, \\ p(j|t) &= N_j(t)/N(t), \end{aligned}$$

où $\tilde{T}$ est l'ensemble des noeuds terminaux, $p(t)$ est la probabilité que le cas soit dans le noeud t, $N(t)$ est le nombre d'observations dans le noeud t, $N_j(t)$ est le nombre d'observations de la classe j dans le noeud t et N est le nombre total d'observations. Aussi, en choisissant la fonction $O\left(\tilde{T}\right) = \alpha \times \text{card}(\tilde{T})$, où α est une constante et $\text{card}(\tilde{T})$ représente la cardinalité de l'ensemble $\tilde{T}$, il existe un arbre qui minimise l'équation (3.4.1), (pour la démonstration, voir Breiman *et al.*, 1984).

3.4.3.1. *Construction d'une suite d'arbres*

Tout d'abord, il faut construire la suite d'arbres suivante :

$$T_{\max} \succ T_1 \succ T_2 \ldots \succ \{t_1\},$$

où $T_{\max}$ est l'arbre de classification avec un maximum de noeuds terminaux et $\{t_1\}$ est l'absence d'arbre. À la limite, chaque cas possède son noeud terminal pour une mauvaise classification nulle ($R(T) = 0$). Nous avons déjà trouvé l'arbre $T_{\max}$ par la section 3.3, alors, il faut maitenant trouver T_1 :

(1) Choisir t_L et t_R deux noeuds terminaux provenant d'un parent t dans l'arbre $T_{\max}$, (L gauche, R : droite).

(2) Si $R(\{t\}) = R(\{t_L\}) + R(\{t_R\})$, alors il faut retirer ces deux branches de l'arbre.

(3) Reprendre 1 et 2 jusqu'à ce ne soit plus possible.

L'arbre résultant sera noté T_1. Ensuite, en partant de l'arbre T_1, nous allons procéder par la méthode « weakest-link cutting » ou le retirement de la branche la moins importante.

Proposition

Pour tout t, noeud non terminal de l'arbre T_1,

$$R_O(\{t\}) > R_O(T_t), \tag{3.4.2}$$

où T_t est une branche de l'arbre T avec comme noeud de départ t et tous ses descendants. Nous avons vu, par l'équation (3.4.1), que

$$R_O(\{t\}) = R(\{t\}) + \alpha, \tag{3.4.3}$$

$$R_O(T_t) = R(T_t) + \alpha(\tilde{T}_t). \tag{3.4.4}$$

Ce qui implique que, en remplaçant (3.4.3) et (3.4.4) dans (3.4.2), nous obtenons l'équation :

$$\alpha < \frac{R(\{t\}) - R(T_t)}{\text{card}(\tilde{T}_t) - 1}.$$

Ainsi, à l'aide de cette proposition, voici les trois autres étapes à suivre :

(1) La valeur de α_1 est $\alpha_1 = 0$.

(2) À l'aide de la fonction $g_1(\{t\}) = \begin{cases} \frac{R(\{t\}) - R(T_t)}{\text{card}(\tilde{T}_t) - 1}, & \{t\} \in \tilde{T}_1, \\ +\infty, & \{t\} \notin \tilde{T}_1, \end{cases}$ trouver la branche la plus faible définie par $\{\bar{t_1}\} = \arg\min_{\{t\} \in (T_1)} g_1(\{t\})$ et $\alpha_2 = g_1(\{\bar{t_1}\})$.

(3) Définir $T_2 = T_1 - T_{\{\bar{t_1}\}}$ (retirer la branche $\{\bar{t_1}\}$ de l'arbre T_1 et appeler le nouvel arbre T_2) et reprendre à l'étape 2 jusqu'à obtenir les séries $T_{\max} \succ T_1 \succ T_2 \ldots \succ \{\{t_1\}\}$, et $\alpha_1 < \alpha_2 < \ldots < \alpha_k < \ldots, k \geqslant 1$.

Le problème est maintenant de déterminer lequel de ces arbres est le meilleur.

3.4.3.2. *Échantillon test*

Pour trouver le choix optimal, nous pouvons utiliser l'estimation par échantillon test. L'estimation par échantillon test consiste en fait à séparer l'échantillon d'apprentissage en deux, un échantillon de construction et un échantillon test. Cette méthode est efficace au niveau du calcul et elle est préférable lorsque l'échantillon d'apprentissage contient un grand nombre d'observations.

(1) Fixer un nombre $N(2)$ d'observations au hasard de l'échantillon L pour former le nouvel échantillon L_2, $L_1 = L - L_2$ est l'échantillon de construction.

(2) Construire la série d'arbres $T_{\max} \succ T_1 \succ T_2 \ldots \succ \{\{t_1\}\}$ par la méthode de la section 3.4.3.1 énoncée précédemment basé sur L_1.

(3) Calculer le taux de mauvaise classification en utilisant l'échantillon L_2 pour chacun des arbres :

$$R^{ts}(T) = \frac{1}{N(2)} \sum_{i,j} c(i|j) N_{ij}(2),$$

où

N_{ij} : nombre d'observations dans la classe j de L_2 classé dans la classe i, et

$c(i|j)$: est le coût de mauvaise classification d'un cas de la classe j dans la classe i.

(4) L'arbre de classification est choisi par $R^{ts}(T_{k_0}) = \min_k R^{ts}(T_k)$.

Bref, le meilleur arbre de la série $T_{\max} \succ T_1 \succ T_2 \ldots \succ \{t_1\}$ avec l'estimation par échantillon test est $R^{ts}(T_{k_0})$. Une méthode alternative est la validation croisée.

3.4.3.3. *Validation croisée*

Cette méthode exige beaucoup de calculs, mais donne une bonne idée de la stabilité de l'arbre de classification. Voici les étapes à suivre :

(1) Diviser l'échantillon L en V parties égales choisies au hasard (souvent $V = 10\%$ de l'échantillon).

(2) Construire l'échantillon v comme étant $L_{L-v} = L - L_v, v = 1, \ldots, V$.

(3) Construire les arbres $T^{(v)}_{\max}, v = 1, \ldots, V$, (le plus de noeuds possibles à partir de l'échantillon L_{L-v}.

(4) Calculer

$$R^{cv}(T) = \frac{1}{N} \sum_{i,j} c(i|j) N_{ij},$$

où $N_{ij} = \sum_v N^{(v)}_{ij}$ est le nombre total d'observations de la classe j classées en i et $N^{(v)}_{ij}$ est le nombre d'observations de la classe j classées en i par l'arbre $T^{(v)}$.

(5) L'arbre de classification est choisi par $R^{cv}(T_{k_0}) = \min_k R^{cv}(T_k)$.

Alors, le meilleur arbre avec la méthode de validation croisée est $R^{cv}(T_{k_0})$. Bref, cette méthode consiste, au lieu de construire un arbre et de retirer des branches, à construire plusieurs petits arbres et à choisir le meilleur parmi ceux-ci.

3.5. VALEURS MANQUANTES, BREIMAN *et al.* (1984)

Dans le cas où il y aurait présence de valeurs manquantes, il s'agit de remplacer la ou les règles de séparation qui ne peuvent être évaluées par leur règle de séparation la plus proche. Pour plus de détails sur les méthodes utilisées lors de présence de valeurs manquantes dans un arbre de décision, voir Breiman *et al.* (1984) ou Ragel (1998).

En conclusion, nous avons vu, dans ce chapitre, la méthodologie à suivre pour construire un arbre de classification en utilisant les observations disponibles. Cette méthodologie est composée d'une règle de séparation (Gini, Twoing ou autres), d'une fontion d'impureté et d'un choix d'un nombre de noeuds terminaux.

Nous voulons maintenant construire l'arbre en incorporant l'information *a priori* que nous avons trouvée au chapitre 2, soit le lien entre la valeur du fonds

et la délinquance. Ainsi, dans le chapitre suivant, nous présentons l'arbre de classification bayésien.

Chapitre 4

ARBRE DE CLASSIFICATION BAYÉSIEN

L'arbre de classification est un système hiérarchique constitué de noeuds, de branches et de noeuds terminaux. La classe associée à ces noeuds terminaux est déterminée par la prévalence des points s'y retrouvant. Dans un modèle bayésien, l'arbre de classification est représenté par un modèle avec une distribution *a posteriori*. Breiman *et al.* (1984) propose une généralisation de l'arbre nécessaire à l'évaluation de la distribution *a posteriori* et, pour que cette méthode soit réalisable, Chipman *et al.* (1998) propose l'utilisation de la méthode de Monte Carlo markovienne à sauts réversibles (« Reversible Jump Monte Carlo Markov Chain » RJ MCMC). Ainsi, les deux composantes principales de cette approche consistent en la spécification d'information *a priori* sur les différentes parties de l'arbre et l'utilisation d'un processus stochastique pour résoudre le système. L'information *a priori* nous permet de diriger le processus stochastique à travers les arbres les plus probables en réduisant le nombre d'itérations nécessaires. Alors, l'approche bayésienne consiste à trouver, à l'aide de ces deux composantes, l'arbre avec le plus grand pouvoir prédictif sur notre variable retard.

Dans ce chapitre, nous commençons par établir les paramètres utilisés pour décrire l'espace de l'arbre de classification. Ensuite, nous définissons la densité *a posteriori* pour un modèle bayésien. Nous énonçons les différentes fonctions et probabilités utilisées dans le calcul de cette densité. Nous décrivons l'algorithme à suivre pour intégrer la densité *a posteriori*, soit les chaînes de Markov Monte Carlo à sauts réversibles. Finalement, nous énonçons une méthode pour l'assignation des

classes à chacun des noeuds et une méthode pour déterminer le meilleur arbre parmi l'espace visité.

4.1. L'ESPACE D'UN ARBRE DE CLASSIFICATION

Un arbre de classification peut être représenté d'une façon unique par le vecteur suivant :

$$\vec{T} = (s_1^{pos}, s_1^{var}, s_1^{regle}; s_2^{pos}, \ldots, s_{q-1}^{regle}), \tag{4.1.1}$$

où q est le nombre de noeuds terminaux de l'arbre et $s_i^{pos}, s_i^{var}, s_i^{regle}$ définissent respectivement la position du noeud dans l'arbre, la variable utilisée comme séparation à ce noeud et la valeur de la séparation. Par définition, la position du premier noeud est $s_1^{pos} = 1$. Les positions des noeuds descendants sont définies par $s_i = 2s_i^{parent}$ si la règle de séparations du noeud parent est respectée sinon, la position est $s_i = 2s_i^{parent} + 1$. Ensuite, la variable s_i^{var} peut prendre les valeurs 1 à M, où M est le nombre de variables explicatives et la variable s_i^{regle} appartient à l'ensemble $\{x_j^{(1)}, \ldots, x_j^{(\kappa)}\}$ où κ est le nombre total de valeurs possibles de séparation pour la variable explicative x_j. Bref, les deux variables s_i^{var} et s_i^{regle} représentent les règles du type $x_j \leqslant x_j^{(l)}, l = 1, 2, \ldots, \kappa$.

Par un exemple, il est facile d'illustrer l'utilisation de ce vecteur. Pour l'arbre de la figure 4.1, $\vec{T} = (s_1^{pos}, s_1^{var}, s_1^{regle}; s_2^{pos}, s_2^{var}, s_2^{regle})$ s'écrit de la façon suivante $\vec{T} = (1, 2, 8, 2, 1, 5)$. Dans cette situation, le paramètre $q = 3$ puisqu'il y a trois noeuds terminaux, ce qui implique que $i \in \{1, 2\}$. Il n'est pas nécessaire de mettre les noeuds terminaux dans le vecteur puisqu'ils sont complètement déterminés par la structure de l'arbre. Aussi, nous remarquons que si le nombre de noeuds terminaux changent, la dimension du modèle change aussi. Maintenant que l'arbre est représenté par un vecteur, définissons notre densité *a posteriori*.

4.2. LA DENSITÉ *a posteriori*

En général, lorsque nous pouvons supposer que les données sont indépendantes et identiquement distribuées, la fonction de vraisemblance est écrite comme un

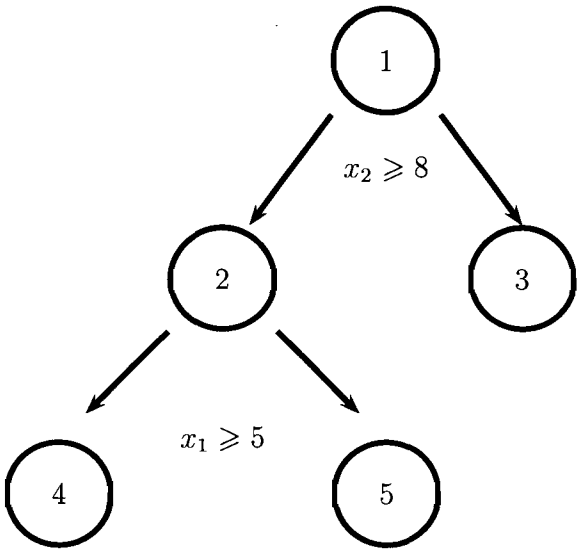


FIGURE. 4.1. Arbre de classification

produit des fonctions de densités :

$$\vec{y} \sim f(\vec{y}|\vec{\zeta}), \qquad f(\vec{y}|\vec{\zeta}) = \prod_{i=1}^{n} f(y_i|\vec{\zeta}),$$

où n est le nombre d'observations et $\vec{\zeta}$ est le vecteur des paramètres décrivant la densité des données. Alors, pour trouver la densité *a posteriori*, nous utilisons l'équation suivante :

$$\pi(\vec{\zeta}|\vec{y}) = \frac{f(\vec{y}|\vec{\zeta}) \times \pi(\vec{\zeta})}{\int f(\vec{y}|\vec{\zeta}) \times \pi(\vec{\zeta}) d\vec{\zeta}},$$

où $\pi(\vec{\zeta})$ est la fonction de densité *a priori*.

Pour trouver la densité *a posteriori* de notre arbre de classification, nous définissons $\vec{y} \sim \Pr(\vec{y}|\vec{x}, \vec{\theta}, \vec{T})$, où $y \in \{1, \ldots, C\}$ est la variable dépendante avec C étant le nombre de classes. Alors, nous obtenons une densité marginale par l'intégrale sur les paramètres de $\vec{\theta} = (\theta_1, \ldots, \theta_q)$ où $\vec{\theta}$ assigne une classe à chacun des noeuds terminaux :

$$\Pr(\vec{y}|\vec{x}, \vec{T}) = \int_{\theta} \Pr(\vec{y}|\vec{x}, \vec{\theta}, \vec{T}) \Pr(\vec{\theta}|\vec{T}) d\vec{\theta}, \tag{4.2.1}$$

où $\vec{x} = (x_1, \ldots, x_M)$ est le vecteur des variables explicatives et $\vec{T}$ est l'arbre de classification. Ensuite, nous pouvons trouver la densité *a posteriori* de l'arbre de classification à une constante près en utilisant l'équation :

$$\Pr(\vec{T}|\vec{x}, \vec{y}) \propto \Pr(\vec{y}|\vec{x}, \vec{T}) \Pr(\vec{T}). \tag{4.2.2}$$

Malheureusement, l'évaluation de (4.2.2) sur tous les arbres possibles ne se calcule pas facilement à l'exception de cas très simple où il existe seulement un petit nombre d'arbres. Ainsi, nous ne pouvons pas trouver la constante de normalisation, et nous ne pouvons pas déterminer quels arbres possèdent la probabilité *a posteriori* la plus élevée. Pour contrer ce problème de constante de normalisation, nous explorons la densité *a posteriori* à l'aide d'un algorithme de type Metropolis-Hastings (voir la section 4.4).

Notre densité *a posteriori* est composée, à une constante près, de deux parties, $\Pr(\vec{y}|\vec{x}, \vec{T})$ et $\Pr(\vec{T})$. Alors, nous définissons ces différentes parties.

4.3. La probabilité $\Pr(\vec{y}|\vec{x}, \vec{T})$

Pour calculer la probabilité $\Pr(\vec{y}|\vec{x}, \vec{T})$,nous avons besoin de trois fonctions :

(1) $\Pr(\vec{y}|\vec{x}, \vec{\theta}, \vec{T})$,

(2) $f(\vec{y}|\vec{\theta})$,

(3) $\Pr(\vec{\theta}|\vec{T})$,

pour pouvoir calculer l'intégrale de l'équation (4.2.1). Ces trois fonctions sont définies comme suit.

4.3.1. La probabilité $\Pr(\vec{y}|\vec{x}, \vec{\theta}, \vec{T})$

Pour les modèles d'arbres de classification, il est supposé que, conditionnellement à $(\vec{\theta}, \vec{T})$, les observations de la variable dépendante sont, dans chacun des noeuds terminaux, indépendantes et identiquement distribuées. De plus, elles sont indépendantes à travers les noeuds terminaux. Si ces hypothèses sont vérifées,

nous pouvons écrire l'équation suivante :

$$\Pr(\vec{y}|\vec{x},\vec{\theta},\vec{T}) = \prod_{i=1}^{q} f(y_i|\theta_i) = \prod_{i=1}^{q}\prod_{j=1}^{n_i} f(y_{ij}|\theta_i), \tag{4.3.1}$$

où y_{ij} est l'observation j dans le noeud i et n_i est le nombre d'observations dans ce noeud.

4.3.2. La fonction $f(\vec{y}|\vec{\theta})$

Dans un arbre de classification à C classes, chacune de ces classes possède une probabilité d'être tirée. Alors, nous pouvons définir la distribution des classes comme une distribution multinomiale.

Définition 4.3.1 (La distribution multinomiale). *Si C et m sont des entiers positifs et si $z_1, \ldots, z_C$ sont des nombres satisfaisant les conditions $0 \leqslant z_j \leqslant 1$, $j = 1, \ldots, C$, et $\sum_{j=1}^{C} z_j = 1$, alors, le vecteur de variables aléatoires $(W_1, \ldots, W_C)$ (compte le nombre de tirages dans chacune des classes) a une distribution multinomiale avec m tirages :*

$$f(w_1, \ldots, w_C) = \frac{m!}{w_1!, \ldots, w_C!} z_1^{w_1}, \ldots, z_n^{w_n} = m! \prod_{j=1}^{C} \frac{z_j^{w_j}}{w_j!}, \tag{4.3.2}$$

où chaque w_i est un entier non négatif et $\sum_{i=1}^{C} w_j = m$.

Nous pouvons réécrire la fonction de densité (4.3.2) pour chacune des observations d'un noeud sous la forme :

$$f(y_{ij}|\theta_i) = \prod_{k=1}^{C} z_{ik}^{I(y_{ij} \in \text{Classe } k)},$$

où $I(y_{ij} \in \text{Classe } k)$ prend la valeur un si le cas de ce noeud est de la classe k, zéro sinon.

4.3.3. La probabilité $\Pr(\vec{\theta}|\vec{T})$

Une façon simple de définir la fonction de densité *a priori* des paramètres z_{ik} d'une distribution multinomiale sur chacun des noeuds est l'utilisation de sa

densité *a priori* conjuguée, soit la distribution de Dirichlet.

Définition 4.3.2 (La fonction de densité Dirichlet). *La fonction de densité Dirichlet avec les paramètres* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_C)$, $\alpha_c > 0$ *où* C *est le nombre de classes est définie par*

$$Dirichlet((z_1, \ldots, z_C)|\vec{\alpha}) = \frac{\Gamma(\sum_{i=1}^{C} \alpha_i)}{\prod_{i=1}^{C} \Gamma(\alpha_i)} \prod_{i=1}^{C} z_i^{\alpha_i - 1},$$

où $z_1, \ldots, z_C \geqslant 0$, $\sum_{i=1}^{C} z_i = 1$ *et* $\alpha_1, \ldots, \alpha_C \geqslant 0$. *Le paramètre* α_i *est interprété comme le nombre d'observations a priori dans une classe. Nous remarquons que si le nombre de classes est deux* ($C = 2$), *cette fonction devient une fonction de densité beta.*

Alors, nous allons utiliser comme densité *a priori* une densité de Dirichlet de la forme :

$$\text{Dirichlet}((z_1, \ldots, z_C)|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{C} \alpha_k)}{\prod_{k=1}^{C} \Gamma(\alpha_k)} \prod_{k=1}^{C} z_{ik}^{\alpha_k - 1}. \tag{4.3.3}$$

Le choix le plus naturel pour le vecteur $\vec{\alpha}$ est $(1, \ldots, 1)$ pour que la fonction de densité *a priori* (4.3.3) soit uniforme sur le simplexe. Par contre, le vecteur $\vec{\alpha}$ peut prendre de plus grandes valeurs pour certaines classes, ayant pour conséquence de rendre ces classes plus sensibles à la mauvaise classification. Ceci est intéressant quand une ou plusieurs classes semblent plus importantes que les autres.

Bref, la fonction $\Pr(\vec{y}|\vec{x}, \vec{T})$ consiste à multiplier la fonction de densité *a priori* de ces paramètres (équation (4.3.3)) par la densité de chacune des observations (équation (4.3.1)) et ce, pour chacun des q noeuds terminaux. Alors, le résultat

est :

$$\begin{aligned}
\Pr(\vec{y}|\vec{x},\vec{T}) &= \int \prod_{i=1}^{q} \left[\text{Dirichlet}((z_1,\ldots,z_C)|\vec{\alpha}) \prod_{j=1}^{n_i} f(y_{ij}|\theta_i) \right] d\vec{z}, \\
&= \int \prod_{i=1}^{q} \left[\text{Dirichlet}((z_1,\ldots,z_C)|\vec{\alpha}) \prod_{j=1}^{n_i} \prod_{k=1}^{C} z_{ik}^{I(y_{ij}\in \text{Classe } k)} \right] d\vec{z}, \\
&= \int \prod_{i=1}^{q} \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \prod_{k=1}^{C} z_{ik}^{\alpha_k-1} \prod_{j=1}^{n_i} \prod_{k=1}^{C} z_{ik}^{I(y_{ij}\in \text{Classe } k)} \right] d\vec{z}, \\
&= \int \prod_{i=1}^{q} \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \prod_{k=1}^{C} z_{ik}^{\alpha_k-1} \prod_{k=1}^{C} z_{ik}^{\sum_{j=1}^{n_i} I(y_{ij}\in \text{Classe } k)} \right] d\vec{z}, \\
&= \int \prod_{i=1}^{q} \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \prod_{k=1}^{C} z_{ik}^{\alpha_k-1} \prod_{k=1}^{C} z_{ik}^{n_{ik}} \right] d\vec{z},
\end{aligned}$$

où $n_{ik} = \sum_{j=1}^{n_i} I(y_{ij} \in \text{Classe } k), n_i = \sum_{k=1}^{C} n_{ik}$, n_{ik} est le nombre d'observations par noeuds terminaux de classe k, $k = 1,\ldots,C$ et n_i est le nombre d'observations dans le noeud i. Ainsi, nous obtenons la probabilité suivante :

$$\Pr(\vec{y}|\vec{x},\vec{T}) = \prod_{i=1}^{q} \int \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \prod_{k=1}^{C} z_{ik}^{\alpha_k+n_{ik}-1} \right] d\vec{z}. \tag{4.3.4}$$

Pour intégrer l'équation (4.3.4), nous remarquons qu'elle suit une densité de Dirichlet de paramètres $\alpha_k + n_{ik}$, (évidemment, nous avons choisi son conjugué comme densité *a priori*). Alors, nous complétons l'intégrale par la constante manquante et nous obtenons :

$$\begin{aligned}
\Pr(\vec{y}|\vec{x},\vec{T}) &= \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \right]^q \prod_{i=1}^{q} \int \left[\frac{\frac{\Gamma(\sum_{k=1}^{C}\alpha_k+n_{ik})}{\prod_{k=1}^{C}\Gamma(\alpha_k+n_{ik})}}{\frac{\Gamma(\sum_{k=1}^{C}\alpha_k+n_{ik})}{\prod_{k=1}^{C}\Gamma(\alpha_k+n_{ik})}} \prod_{k=1}^{C} z_{ik}^{\alpha_k+n_{ik}-1} \right] d\vec{z}, \\
&= \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \right]^q \prod_{i=1}^{q} \frac{\prod_{k=1}^{C}\Gamma(\alpha_k+n_{ik})}{\Gamma(\sum_{k=1}^{C}\alpha_k+n_{ik})} \times \\
&\qquad \int \left[\frac{\Gamma(\sum_{k=1}^{C}\alpha_k+n_{ik})}{\prod_{k=1}^{C}\Gamma(\alpha_k+n_{ik})} \prod_{k=1}^{C} z_{ik}^{\alpha_k+n_{ik}-1} \right] d\vec{z}, \\
&= \left(\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)} \right)^q \prod_{i=1}^{q} \left[\frac{\prod_{k=1}^{C}\Gamma(\alpha_k+n_{ik})}{\Gamma(n_i+\sum_{k=1}^{C}\alpha_k)} \right].
\end{aligned}$$

Maintenant que nous avons trouvé $\Pr(\vec{y}|\vec{x},\vec{T})$, il faut calculer la valeur de $\Pr(\vec{T})$.

4.3.4. La probabilité $\Pr(\vec{T})$

La probabilité d'obtenir un arbre en particulier, $\Pr(\vec{T})$, est déterminée par quatre densités *a priori*, soit celle :

(1) pour la sélection d'une variable explicative associée à ce noeud,

(2) pour la sélection d'une règle de séparation,

(3) pour la forme de l'arbre,

(4) pour le nombre de noeuds terminaux.

Ces quatre densités peuvent être combinées ensemble pour donner la probabilité de choisir l'arbre T avec q noeuds terminaux :

$$\Pr(\vec{T}) = \left\{ \prod_{i=1}^{q-1} \Pr(s_i^{regle} | s_i^{var}) \Pr(s_i^{var}) \right\} \Pr(\{s_i^{pos}\}_1^{q-1}) \ \Pr(q \text{ noeuds terminaux}). \tag{4.3.5}$$

En d'autres termes, la probabilité d'obtenir un arbre en particulier est trouvée par une multiplication des probabilités pour chacun des $q-1$ noeuds non terminaux. Ainsi, il faut multiplier ensemble la probabilité d'avoir choisi une règle de séparation sachant la variable explicative fois la probabilité d'obtenir cette variable, ceci est répété pour tous les noeuds intermédiaires. Ensuite, ce résultat doit être multiplié par la probabilité d'obtenir cette forme d'arbre en particulier et par la probabilité d'obtenir ce nombre de noeuds terminaux.

Pour calculer la probabilité $\Pr(\vec{T})$, il faut donc définir ces quatre densités *a priori*.

4.3.4.1. *Densité a priori sur la sélection d'une variable explicative*

Une fonction de densité *a priori* simple est l'utilisation de la fonction de densité uniforme. Elle permet de choisir les variables explicatives avec la même probabilité. Cette fonction de densité est donnée par :

$$\Pr(s_i^{var}) = \frac{1}{M}.$$

où s_i représente une variable explicative et M est le nombre de variables explicatives.

De plus, une densité avec des probabilités différentes pour chacune des variables peut être intéressante à envisager. Par exemple, nous pourrions préférer les modèles avec peu de variables et mettre une densité qui augmente la probabilité qu'une variable soit choisie si elle a déjà été choisie. Encore, nous pourrions tout simplement accorder à certaines variables que nous jugeons plus importantes, un poids supérieur aux autres.

Alors, nous proposons, dans ce mémoire, comme solution alternative à l'uniforme, une densité basée sur le résultat du test de Wilcoxon. Le test de Wilcoxon calcule la probabilité que l'hypothèse nulle soit vérifiée, c'est-à-dire que les clients sans retard aient la même moyenne que les clients avec des retards (voir la section 1.3 pour le test de Wilcoxon). Ainsi, en utilisant l'équation suivante :

$$\Pr(s_i^{var}) = \frac{1 - \text{valeur-p } (s_i^{var})}{\sum_{j=1}^{M}(1 - \text{valeur-p } (x_j))},$$

où la valeur-p (x_j) provient du test de Wilcoxon, nous obtenons une densité pondérée par la valeur-p du test utilisé. Plus la valeur-p se rapproche de 1, plus le poids de cette variable tend vers 0 et plus la valeur-p se rapproche de 0, plus le poids tend vers 1. Bref, si toutes les variables ont la même valeur-p, nous retrouvons la fonction de densité uniforme.

Nous avons choisi le test de Wilcoxon, car nous ne voulons pas que le test soit influencé par les données éloignées de la moyenne, d'où l'utilisation d'un test non paramétrique. Aussi, nous avons choisi cette densité, car plus une variable explicative discrimine les clients sans retard de ceux ayant des retards, plus celle-ci possède une probabilité élevée d'être choisie et vice-versa.

4.3.4.2. *Densité a priori sur la sélection d'une valeur de séparation*

Après avoir choisi la variable explicative du noeud, il faut maintenant choisir sa règle de séparation. Ainsi, la règle de séparation est conditionnelle à la variable

explicative x_i. À cause d'un problème pratique, nous pouvons considérer que l'espace possible des règles de séparation est un ensemble discret et fini. Nous ne pouvons envisager ceci comme une restriction puisque la base de données contient nécessairement un nombre de cas fini. Comme pour le choix de la variable explicative, une fonction de densité *a priori* simple est l'utilisation de la fonction de densité uniforme. Voici la fonction de densité :

$$\Pr(s_i^{regle}|s_i^{var}) = \frac{1}{\kappa_i},$$

où κ_i représente le nombre de valeurs différentes de la variable explicative s_i^{var}. Cette probabilité n'est pas influencée par les répétitions, c'est-à-dire qu'une valeur répétée plusieurs fois ne possède pas une probabilité plus élevée d'être sélectionnée que les autres. Malgré que ce choix de densité soit simple, il faut remarquer que la probabilité $1/\kappa_i$ décroît avec le nombre de règles de séparation possibles. Donc, plus une variable explicative a de valeurs différentes, plus la probabilité de choisir une règle spécifique diminue. Cette propriété est aussi appelée l'effet de dilution et elle est importante lors du choix du meilleur arbre de classification. Un effet de dilution se définit comme ceci.

Définition 4.3.3 (Effet de dilution). *Un effet de dilution est une diminution de la probabilité de choisir un élément lorsque la cardinalité de l'espace où l'élément est choisi augmente. Par exemple, une variable continue possède une probabilité plus faible d'être choisie qu'une variable discrète, (voir Chipman et al. (2001) pour plus de détails sur les effets de dilutions).*

Donc, il faut se demander si cette situation reflète la réalité. Par exemple, est-ce qu'une variable continue possède une probabilité plus faible de prédire la présence d'un retard sur le remboursement d'un prêt qu'une variable discrète ?

Une alternative à ce problème est la pondération de la fonction de densité du choix de la variable par un facteur pour venir compenser l'augmentation ou la diminution de κ. Par contre, il est difficile de justifier que certaines variables explicatives ont une probabilité plus élevée d'être sélectionnées que d'autres sans

information *a priori*. De plus, avec cette alternative, nous avons uniquement déplacé le problème au choix de la variable.

Dans l'article de Chipman *et al.* (1998), les auteurs ont essayé plusieurs densités afin de corriger ce problème. Tout d'abord, une première méthode définit les bornes de la fonction de densité uniforme comme la distance entre le minimum et le maximum plûtot que sur la fréquence. Ceci engendre quand même un problème du fait qu'une variable avec une grande étendue possède des règles de séparation avec une probabilité plus faible d'être choisie qu'une variable moins étendue (problème d'échelle). Une deuxième méthode suggère une fonction qui accorde moins d'importance aux petites et grandes valeurs, car nous nous attendons peut-être à ce qu'il y ait plus de probabilité qu'une séparation advienne au milieu de la distribution.

Dans ce mémoire, nous proposons deux fonctions de densité *a priori* que nous allons confronter dans la section des résultats. Les deux fonctions n'ont été testées que pour le cas de deux classes et elles devront être modifiées advenant une application à plusieurs classes. La première consiste en une fonction de densité de Cauchy.

Définition 4.3.4 (Densité de Cauchy).

$$\begin{aligned} f(s|\mu,\sigma) &= \frac{1}{\pi\gamma\left[1+\left(\frac{s-\mu}{\gamma}\right)^2\right]}, \qquad s,\mu \in \infty, \gamma > 0, \\ F(s|\mu,\sigma) &= \frac{1}{\pi}\arctan\left(\frac{s-\mu}{\gamma}\right)+\frac{1}{2}. \end{aligned}$$

L'estimation des paramètres de position et d'échelle de cette densité *a priori* se fait de la manière suivante :

(1) trouver la fonction de répartition empirique de chacune des deux classes et repérer l'endroit où l'écart est maximal,

(2) fixer le paramètre de location μ à cette valeur,

(3) isoler le paramètre d'échelle γ de la fonction de distribution avec 95 % de la densité comprise entre l'intervalle (minimum, 2μ − minimum) ou (2μ − maximum, maximum). Nous choisissons, parmi ces deux intervalles, le plus grand.

Voici l'équation de distribution avec le paramètre d'échelle γ isolé :

$$\begin{aligned}
\gamma &= \frac{1}{2(-1+a)}(-1i\varphi - (-\varphi^2 - 4(-1+a) \times \\
&\quad (-\mu^2 + a\mu^2 + \mu\psi_2 - a\mu\psi_2 + \mu\psi_1 - a\mu\psi_1 - \psi_1\psi_2 + a\psi_1\psi_2))^{0.5}), \\
a &= e^{2i0,95\pi}, \\
\varphi &= -\psi_2 - a\psi_2 + \psi_1 + a\psi_1,
\end{aligned}$$

où $i = \sqrt{-1}$ et ψ_1 = minimum et ψ_2 = maximum correspondent aux extremum de l'intervalle spécifié en (3), car nous voulons obtenir une probabilité non nulle pour chacune des valeurs de séparation possibles. Ainsi, cette fonction de densité donne une légère importance (la densité Cauchy possède des ailes relevées) aux règles de séparation se trouvant à l'endroit où l'écart entre les deux fonctions de répartition est maximal.

Contrairement à la fonction de densité uniforme, les probabilités ne sont pas influencées par la fréquence, mais plutôt par le paramètre d'échelle. Ainsi, plus le paramètre d'échelle est élevé pour une variable explicative, plus la probabilité de choisir une règle de séparation en particulier diminue.

En plus d'avoir testé cette méthode sur notre échantillon, nous avons modifié la probabilité de choisir les variables explicatives de façon à ce que le paramètre d'échelle n'impacte plus la probabilité de choisir la règle de séparation. Voici un résumé des fonctions de densité *a priori* testées sur notre échantillon pour le choix de la variable explicative et de sa règle de séparation :

Situation 1 :

$$\begin{aligned}\Pr(s_i^{var}) &= \frac{1-\text{valeur-p }(s_i^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))},\\ \Pr(s_i^{regle}|s_i^{var}) &= \frac{1}{\pi\gamma_i\left[1+\left(\frac{s_i^{regle}-\mu_i}{\gamma_i}\right)^2\right]}.\end{aligned}$$

Situation 2 :

$$\begin{aligned}\Pr(s_i^{var}) &= \frac{\gamma_{s_i^{var}}}{\sum_{i=1}^{M}\gamma_{x_i}},\\ \Pr(s_i^{regle}|s_i^{var}) &= \frac{1}{\pi\gamma_i\left[1+\left(\frac{s_i^{regle}-\mu_i}{\gamma_i}\right)^2\right]}.\end{aligned}$$

Une autre alternative est la fonction de densité construite à partir du D de Somers (voir 5.3 pour les détails du D de Somers). Le D de Somers est une mesure d'association allant de -1 dans la cas d'une discordance parfaite à 1, une concordance parfaite. Voici les fonctions de densité utilisées :

$$\begin{aligned}\Pr(s_i^{var}) &= \frac{1-\text{valeur-p }(s_i^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))},\\ \Pr(s_i^{regle}|s_i^{var}) &\propto \max_{j\in\{1,\ldots,\kappa\}}\left(D_j(s_i^{regle}|s_i^{var}),\epsilon\right),\end{aligned}$$

où ϵ est une valeur choisie arbitrairement petite et $D_j(s_i^{regle}|s_i^{var})$ est le D de Somers. La valeur ϵ vient donner une probabilité non nulle aux règles de séparation possédant un D de Somers négatif. Dans ce mémoire, nous proposons comme choix de ϵ la valeur suivante :

$$\epsilon = \min_{j\in\{1,\ldots,\kappa\}}\left(D_j(s_i^{regle}|s_i^{var})^+\right),$$

Ainsi, les règles de séparation possédant une mesure d'association élevée ont une plus grande probabilité d'être choisies que les autres règles. Les inconvénients de ce choix de densité sont l'ampleur des calculs puisqu'il faut calculer pour chacune des valeurs différentes de l'échantillon le D de Somers et, comme pour l'uniforme, la probabilité de choisir une règle de séparation diminue avec le nombre de valeurs possibles.

4.3.4.3. *Densité a priori sur la forme de l'arbre*

Dans les sections précédentes, nous avons trouvé une loi *a priori* sur la règle et sur la variable. Il faut multiplier ceux-ci par la probabilité d'obtenir cette forme d'arbre, car il existe plusieurs façons, par exemple, de dessiner un arbre à 4 noeuds terminaux (cela varie selon la profondeur de chacun des noeuds terminaux). Alors, la probabilité $\Pr(\{s_i^{pos}\}_1^{q-1})$ représente ces façons de dessiner et cette probabilité se définit comme :

$$\Pr(\{s_i^{pos}\}_1^{q-1}) = \frac{1}{S_q},$$

où S_q est le nombre de Catalan (Campbell, 1984) et $\{\cdot\}_1^{q-1}$ représente les différentes combinaisons. Le nombre de Catalan représente habituellement le nombre de différentes façons de partager un polygone en triangles. Par contre, selon Schetinin (2004), ce nombre représente aussi les différentes façons de dessiner un arbre. Alors, le nombre de Catalan est :

$$S_q = \frac{1}{q+1}\binom{2q}{q}.$$

où q est le nombre de noeuds terminaux.

4.3.4.4. *Densité a priori sur le nombre de noeuds terminaux*

L'avantage d'un arbre de classification bayésien se trouve dans le fait qu'il n'existe pas de problème de complexité puisque celui-ci se règle avec la densité *a priori* sur le nombre de noeuds terminaux. Par contre, il faut connaître, à l'avance, un estimé de la quantité de noeuds terminaux nécessaires à notre arbre de classification. L'article de Chipman *et al.* (1998) propose deux fonctions de densités. La première fonction, la plus simple, est définie comme ceci :

$$\Pr(\eta|\vec{T}) = \alpha, \qquad \text{où } \alpha < 1,$$

et η représente le noeud où la décision de le déclarer terminal ou non est prise. Sous cette densité, la probabilité d'obtenir un arbre de q noeuds terminaux suit une loi géométrique de la forme :

$$\Pr(q \text{ noeuds terminaux}) = \left\{\frac{1-\alpha+\alpha^2}{1-\alpha}\right\}\alpha^{q-1}(1-\alpha)^q.$$

L'inconvénient d'une loi géométrique est que cette fonction ne tient pas compte de la profondeur des noeuds. Ainsi, elle accorde une probabilité égale pour chacun des arbres avec q noeuds terminaux, peu importe leur forme.

La deuxième méthode vise à corriger ce problème en introduisant un facteur de profondeur :

$$\Pr(\eta|\vec{T}) = \alpha(1+d_\eta)^{-\beta}, \qquad \text{où } \alpha < 1, \beta \geqslant 0,$$

et d_η mesure la profondeur du noeud η. Cette fonction décroît avec d_η selon la valeur de β. Ainsi, nous pouvons contrôler le nombre de noeuds de l'arbre par α et la profondeur par β. Contrairement à la première méthode, cette méthode donne une probabilité plus élevée aux arbres avec des branches de même longueur. Un inconvénient réside dans le fait que nous devons posséder de l'information *a priori* supplémentaire pour déterminer la valeur du paramètre β.

Dans ce mémoire, nous utilisons plutôt, comme fonction de densité *a priori*, une distribution de Poisson tronquée proposée par Denison *et al.* (2002) :

$$\Pr(q \text{ noeuds terminaux}) = \frac{\lambda^q}{(e^\lambda - 1)q!}, \qquad \text{où } \lambda \geqslant 0.$$

Nous utilisons une distribution tronquée, car nous avons au minimum un noeud terminal dans notre arbre. Ainsi, nous n'agissons pas sur la longueur des branches de l'arbre avec ce choix de densité.

Maintenant que nous avons une fonction de densité *a priori* sur le choix des variables, sur la valeur de séparation, sur la forme de l'arbre, sur le nombre de noeuds terminaux et que nous avons une expression pour $\Pr(\vec{T})$, nous pouvons développer l'algorithme des chaînes de Markov Monte Carlo à sauts réversibles. Cet algorithme nous permet d'évaluer la probabilité de l'équation (4.2.2) sans calculer l'intégrale.

4.4. ALGORITHME : CHAÎNES DE MARKOV MONTE CARLO À SAUTS RÉVERSIBLES

L'algorithme des chaînes de Markov Monte Carlo à sauts réversibles (RJ MCMC) est introduit par Green (1995) et il permet de sauter, comme son nom le dit, à travers les sous-espaces de tailles différentes. Celui-ci augmente considérablement l'étendue d'application de la méthode de Metropolis-Hastings (Metropolis *et al.*, 1953 et Hastings, 1970) et cet algorithme s'applique aux différents problèmes possédant des changements de dimension. En bref, la méthode de Metropolis-Hastings consiste à se promener à travers l'espace, de façon systématique ou aléatoire. Après avoir visité un certains nombre d'états, les paramètres sont mis à jour. Cette mise à jour est acceptée ou refusée selon un ratio de densité ciblée multipliée par un ratio de densité proposée. Cette méthode est répétée jusqu'à la convergence de la distribution qui nous intéresse. Il s'agit d'adapter cet algorithme en fonction d'un changement d'espace, voir Green (1995) pour tous les détails sur les changements d'espace.

Pour les arbres de classification, avec une forme et un nombre de noeuds fixes, nous pouvons faire une marche aléatoire sur les noeuds intérieurs et ainsi, obtenir un système d'une dimension fixe. Dans ces conditions, la méthode de Metropolis-Hastings s'avère suffisante. Par contre, pour construire notre arbre, nous allons rajouter des noeuds et en enlever. Ainsi, la dimension change à chaque naissance ou mort d'un noeud, d'où l'utilisation de la méthode de Green (1995).

Alors, pour contrer le problème de la constante de normalisation de l'équation (4.2.2), nous simulons une série d'arbres avec une chaîne de Markov :

$$\vec{T}^0, \vec{T}^1, \vec{T}^2, \ldots, \tag{4.4.1}$$

où leur distribution converge sous certaines conditions, énoncé par Tierney (1994), vers notre distribution *a posteriori* $\Pr(\vec{T}|\vec{x}, \vec{y})$. Parce que cette série d'arbres se retrouve autour des régions de probabilité *a posteriori* maximale, cette simulation est utilisée pour trouver stochastiquement les arbres de classification avec le

meilleur pouvoir prédictif.

L'algorithme à sauts réversibles pour simuler la chaîne de Markov (4.4.1) est défini comme suit. Il faut donner une valeur initiale à l'arbre $\vec{T}^0$. Par exemple, nous définissons le premier noeud : $\vec{T}^0 = c(s_1^{pos} = 1, s_1^{var}, s_1^{regle})$, et nous simulons, d'une façon itérative, la transition de l'arbre $\vec{T}^i$ à l'arbre $\vec{T}^{i+1}$ par la méthode suivante :

(1) Générer un arbre $\vec{T}^*$ avec la distribution de probabilité $f(\vec{T}^i, \vec{T}^*)$ définie aux sections 4.4.1, 4.4.2 et 4.4.3.

(2) Poser $\vec{T}^{i+1} = \vec{T}^*$ avec la probabilité d'acceptation suivante :

$$\alpha(\vec{T}^i, \vec{T}^*) = \min\left\{\frac{f(\vec{T}^i|\vec{T}^*)}{f(\vec{T}^*|\vec{T}^i)}\frac{\Pr(\vec{y}|\vec{x}, \vec{T}^*)\Pr(\vec{T}^*)}{\Pr(\vec{y}|\vec{x}, \vec{T}^i)\Pr(\vec{T}^i)}, 1\right\} \tag{4.4.2}$$

sinon, poser $\vec{T^{i+1}} = \vec{T}^i$.

Nous remarquons que la constante de normalisation n'est pas nécessaire pour calculer l'équation (4.4.2) puisqu'elle se simplifie par la division.

Alors, pour appliquer cette itération, nous connaissons les deux probabilités $\Pr(\vec{y}|\vec{x}, \vec{T})$ et $\Pr(\vec{T})$ par les sections précédentes. Par conséquent, il faut définir la distribution de probabilité de $f(\vec{T}^i|\vec{T}^*)$ et $f(\vec{T}^*|\vec{T}^i)$. Cette distribution correspond en fait à la probabilité accordée à chacun des mouvements effectués pour explorer la distribution *a posteriori*. Nous conservons les mêmes mouvements énoncés dans Chipman *et al.* (1998) et Denison *et al.* (2002), soit :

(1) **la naissance** : choisir un noeud terminal de façon aléatoire et le séparer en son noeud de droite et de gauche avec une variable explicative et une règle de séparation.

(2) **La mort** : choisir, de façon aléatoire, un noeud parent dont les deux enfants sont des noeuds terminaux et retirer ces enfants afin que ce noeud parent soit maintenant terminal.

(3) **Le changement de la variable explicative** : choisir un noeud non terminal, de façon aléatoire, et changer la variable explicative à ce noeud par une autre. Par le fait même, il faut aussi changer la règle de séparation pour cette nouvelle variable.

(4) **Le changement de la règle de séparation** : choisir un noeud non terminal, de façon aléatoire, et changer la règle de séparation pour une autre règle.

Ces mouvements doivent respecter les fonctions de densités *a priori* définies dans la section précédente. Par exemple, pour une naissance, la variable explicative est choisie avec probabilité $\Pr(s_i^{var})$ et sa règle de séparation avec probabilité $\Pr(s_i^{regle}|s_i^{var})$.

De plus, nous choisissons le mouvement avec les probabilités suivantes :

(1) **la naissance** : $b_i = 2\tau \min\left(1, \frac{\Pr(\eta+1|\vec{T^*})}{\Pr(\eta|\vec{T^i})}\right)$,

(2) **la mort** : $d_i = \tau \min\left(1, \frac{\Pr(\eta|\vec{T^i})}{\Pr(\eta+1|\vec{T^*})}\right)$,

(3) **le changement de la variable explicative** : $v_i = (1 - b_i - d_i)/2$,

(4) **le changement de la règle de séparation** : $r_i = (1 - b_i - d_i)/2$,

où τ est une constante forçant la probabilité :

$$b_i + d_i \leqslant 0{,}75 \qquad \text{où } i = 1, 2, 3, \ldots . \tag{4.4.3}$$

Cette contrainte est nécessaire pour éviter que l'algorithme change trop souvent de dimension avant même de l'avoir explorer. Nous fixons $\tau = 0{,}2$ puisque cette valeur force le respect de la condition (4.4.3) avec comme fonction de probabilité $\Pr(\eta|\vec{T})$, une loi de Poisson. Pour $i = 0$, nous avons $b_0 = 1$ et $d_0 = v_0 = r_0 = 0$. La probabilité de naissance se trouve à être deux fois la probabilité de mort parce que nous ne pouvons avoir de naissance si le noeud est homogène. Alors, nous multiplions par deux la probabilité pour conserver un taux similaire d'acceptation de naissance et de mort, voir Denison et al. (1998). Nous assignons cette probabilité v_i et r_i au changement des variables et des règles pour que le mouvement ait la même probabilité d'être choisie. La raison de ce choix provient du fait que nous

n'avons pas d'information nous indiquant que nous devrions changer la variable plus souvent que la règle ou vice-versa.

Ensuite, avec la description de ces mouvements dans l'espace, nous pouvons maintenant définir les fonctions $f(\vec{T}^i|\vec{T}^*)$ et $f(\vec{T}^i|\vec{T}^*)$, c'est-à-dire le mouvement et son inverse. Ces fonctions sont différentes pour chacun des mouvements puisqu'elles correspondent respectivement à la probabilité que celui-ci arrive versus son mouvement inverse. Voici la définition de la fonction $f(\cdot)$ pour les différents mouvements avec les fonctions *a priori* choisies pour ce mémoire.

4.4.1. La naissance

Dans cette section, nous définissons les différentes fonctions nécessaires pour calculer la probabilité d'acceptation d'une naissance. Nous commençons par définir la probabilité du mouvement.

La fonction du mouvement $f(\vec{T}^*|\vec{T}^i)$ se définit comme suit :

(1) choisir l'étape de la naissance avec probabilité : $2 \times \tau \times \min(1, \lambda/(q+1))$. Cette probabilité se réduit à $0{,}4 \times \min(1, \lambda/(q+1))$ pour une loi de Poisson (section 4.3.4.4) comme densité *a priori* sur le nombre de noeuds terminaux.

(2) Choisir le lieu de la naissance (choisir un noeud terminal) avec probabilité : $1/q$ où q est le nombre de noeuds terminaux.

(3) Choisir une variable s_q^{var} avec probabilité : $\frac{1-\text{valeur-p } (s_q^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p } (x_j))}$ (section 4.3.4.1) ou $\frac{\gamma_{s_q^{var}}}{\sum_{i=1}^{M}\gamma_{x_i}}$ (section 4.3.4.2).

(4) Choisir une règle s_q^{regle} avec probabilité : $\frac{1}{\pi\gamma_q\left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right]}$ (section 4.3.4.2).

La fonction du mouvement inverse $f(\vec{T}^i|\vec{T}^*)$ se définit comme suit :

(1) Choisir la mort avec probabilité : $\tau \times \min(1, (q+1)/\lambda)$ où $\tau = 0{,}2$.

(2) Choisir le noeud parent à retirer avec probabilité : $1/(q_{die}+1)$ où q_{die} correspond au nombre de noeuds parents qui possèdent deux noeuds enfants terminaux.

Nous choisissons $\Pr(s_q^{var}) = \frac{1-\text{valeur-p }(s_q^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))}$ pour fin de calcul. Le ratio de l'équation (4.4.2) devient alors :

$$\begin{aligned} f(\vec{T}^*|\vec{T}^i) &= \Pr(\text{naissance}) \times \Pr(\text{lieu de naissance}) \\ &\quad \times \Pr(\text{choix de la variable}) \times \Pr(\text{choix de la règle}) , \\ &= \frac{2 \times 0,2 \times \min(1, \lambda/(q+1)) \{1 - \text{valeur-p } (s_q^{var})\}}{q\ \pi\gamma_q \left[1 + \left(\frac{s_q^{regle} - \mu_q}{\gamma_q}\right)^2\right] \sum_{j=1}^{M}(1 - \text{valeur-p } (x_j))}, \\ f(\vec{T}^i|\vec{T}^*) &= \Pr(\text{mort}) \times \Pr(\text{lieu de la mort}), \\ &= \frac{0,2 \times \min(1, (q+1)/\lambda)}{(q_{die}+1)}. \end{aligned}$$

Ainsi,

$$\begin{aligned} \frac{f(\vec{T}^i|\vec{T}^*)}{f(\vec{T}^*|\vec{T}^i)} &= \frac{\frac{0,2 \times \min(1,(q+1)/\lambda)}{(q_{die}+1)}}{\frac{2 \times 0,2 \times \min(1,\lambda/(q+1))\{1-\text{valeur-p }(s_q^{var})\}}{q\ \pi\gamma_q \left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right] \sum_{j=1}^{M}(1-\text{valeur-p }(x_j))}}, \\ &= \frac{\min(1, (q+1)/\lambda)}{\min(1, \lambda/(q+1))} \times \\ &\quad \frac{q\pi\gamma_q \left[1 + \left(\frac{s_q^{regle} - \mu_q}{\gamma_q}\right)^2\right] \sum_{j=1}^{M}(1 - \text{valeur-p } (x_j))}{2(q_{die}+1)\{1 - \text{valeur-p } (s_q^{var})\}}, \\ &= \frac{q+1}{\lambda} \frac{q\pi\gamma_q \left[1 + \left(\frac{s_q^{regle} - \mu_q}{\gamma_q}\right)^2\right] \sum_{j=1}^{M}(1 - \text{valeur-p } (x_j))}{2(q_{die}+1)\{1 - \text{valeur-p } (s_q^{var})\}} \end{aligned} \quad (4.4.4)$$

De plus, avec les définitions des lois *a priori*, nous pouvons calculer la valeur du ratio $\frac{\Pr(\vec{T}^*)}{\Pr(\vec{T}^i)}$ en utilisant l'équation (4.3.5) :

$$\begin{aligned} \Pr(\vec{T}, q) &= \left\{\prod_{i=1}^{q-1} \Pr(s_i^{regle}|s_i^{var}) \Pr(s_i^{var})\right\} \\ &\quad \times \Pr(\{s_i^{pos}\}_1^{q-1})\ \Pr(q \text{ noeuds terminaux}), \\ &= \left\{\prod_{i=1}^{q} \frac{1}{\pi\gamma_i \left[1 + \left(\frac{s_i^{regle} - \mu_i}{\gamma_i}\right)^2\right]} \frac{1 - \text{valeur-p } (s_i^{var})}{\sum_{j=1}^{M}(1 - \text{valeur-p } (x_j))}\right\} \\ &\quad \times \frac{\lambda^q}{(e^\lambda - 1)(q)! S_q}, \end{aligned} \quad (4.4.5)$$

$$\begin{aligned}
\frac{\Pr(\vec{T}^*, q)}{\Pr(\vec{T}^i, q+1)} &= \frac{1}{\pi\gamma_q\left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right]} \frac{1-\text{valeur-p }(s_q^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))} \frac{\lambda}{q+1} \frac{S_q}{S_{q+1}}, \\
&= \frac{1-\text{valeur-p }(s_q^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))} \\
&\quad \times \frac{q+2}{\pi\gamma_q\left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right] 2(2q+1)} \frac{\lambda}{q+1}. \qquad (4.4.6)
\end{aligned}$$

Alors, en multipliant l'équation (4.4.4) avec (4.4.6), nous obtenons :

$$\begin{aligned}
\frac{f(\vec{T}^i|\vec{T}^*)}{f(\vec{T}^*|\vec{T}^i)} \frac{\Pr(\vec{T}^*, q)}{\Pr(\vec{T}^i, q+1)} &= \frac{q+1}{\lambda} \frac{\lambda}{(q+1)} \frac{q\pi\gamma_q\left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right]}{2(q_{die}+1)} \\
&\quad \times \frac{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))}{\{1-\text{valeur-p }(s_q^{var})\}} \frac{1-\text{valeur-p }(s_q^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))} \\
&\quad \times \frac{q+2}{\pi\gamma_q\left[1+\left(\frac{s_q^{regle}-\mu_q}{\gamma_q}\right)^2\right] 2(2q+1)}, \\
&= \frac{q(q+2)}{4(2q+1)}. \qquad (4.4.7)
\end{aligned}$$

Nous remarquons par l'équation (4.4.7) que la probabilité d'acceptation (4.4.2) ne dépend pas des fonctions de densités *a priori* choisies puisqu'elles se simplifient avec la probabilité du mouvement. L'équation d'acceptation pour la naissance peut maintenant s'écrire de cette façon :

$$\begin{aligned}
\alpha(\vec{T}^i, \vec{T}^*) &= \min\left\{\frac{f(\vec{T}^i|\vec{T}^*)}{f(\vec{T}^*|\vec{T}^i)} \frac{\Pr(\vec{y}|\vec{x}, \vec{T}^*)\Pr(\vec{T}^*)}{\Pr(\vec{y}|\vec{x}, \vec{T}^i)\Pr(\vec{T}^i)}, 1\right\}, \\
&= \min\left\{\frac{q(q+2)}{4(2q+1)} \frac{\Pr(\vec{y}|\vec{x}, \vec{T}^*)}{\Pr(\vec{y}|\vec{x}, \vec{T}^i)}, 1\right\}, \\
&= \min\left\{\frac{q(q+2)}{4(2q+1)} \frac{\left(\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)}\right)^{q+1} \prod_{i=1}^{q+1}\left[\frac{\prod_{k=1}^{C}\Gamma(n_{ik}+\alpha_k)}{\Gamma(n_i+\sum_{k=1}^{C}\alpha_k)}\right]}{\left(\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)}\right)^{q} \prod_{i=1}^{q}\left[\frac{\prod_{k=1}^{C}\Gamma(n_{ik}+\alpha_k)}{\Gamma(n_i+\sum_{k=1}^{C}\alpha_k)}\right]}, 1\right\}, \\
&= \min\left\{\frac{q(q+2)}{4(2q+1)}\left(\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)}\right)\left[\frac{\prod_{k=1}^{C}\Gamma(n_{(q+1)k}+\alpha_k)}{\Gamma(n_{q+1}+\sum_{k=1}^{C}\alpha_k)}\right], 1\right\}.
\end{aligned}$$

Nous avons trouvé la probabilité d'acceptation pour la naissance. Il faut maintenant trouver cette même probabilité pour les trois autres mouvements.

4.4.2. La mort

Le calcul de la probabilité d'acceptation pour le mouvement de la mort se fait directement puisqu'il correspond à l'inverse de la probabilité de la naissance. Donc, la probabilité d'acceptation pour la mort est :

$$\begin{aligned}\alpha(\vec{T}^*, \vec{T}^i) &= \min\left\{\frac{f(\vec{T}^*|\vec{T}^i)}{f(\vec{T}^i|\vec{T}^*)}\frac{\Pr(\vec{y}|\vec{x},\vec{T}^i)\Pr(\vec{T}^i)}{\Pr(\vec{y}|\vec{x},\vec{T}^*)\Pr(\vec{T}^*)}, 1\right\}, \\ &= \min\left\{\frac{4(2q+1)}{q(q+2)}\left(\frac{\prod_{k=1}^{C}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{C}\alpha_k)}\right)\left[\frac{\Gamma(n_{q+1}+\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(n_{(q+1)k}+\alpha_k)}\right], 1\right\}.\end{aligned}$$

Nous acceptons le mouvement de la mort avec la probabilité $\alpha(\vec{T}^*, \vec{T}^i)$.

4.4.3. Le changement de la variable explicative et de la règle de séparation

Pour le changement de la variable explicative et de la règle de séparation, la probabilité d'acceptation est directe, puisque nous n'avons pas de changement de dimension. Alors, le ratio $\frac{f(\vec{T}^*|\vec{T}^i)}{f(\vec{T}^i|\vec{T}^*)}\frac{\Pr(\vec{T}^i)}{\Pr(\vec{T}^*)}$ est égal à 1 et la probabilité d'acceptation est :

$$\begin{aligned}\alpha(\vec{T}^*, \vec{T}^i) &= \min\left\{\frac{f(\vec{T}^*|\vec{T}^i)}{f(\vec{T}^i|\vec{T}^*)}\frac{\Pr(\vec{y}|\vec{x},\vec{T}^i)\Pr(\vec{T}^i)}{\Pr(\vec{y}|\vec{x},\vec{T}^*)\Pr(\vec{T}^*)}, 1\right\}, \\ &= \min\left\{\left(\frac{\Gamma(\sum_{k=1}^{C}\alpha_k)}{\prod_{k=1}^{C}\Gamma(\alpha_k)}\right)\left[\frac{\prod_{k=1}^{C}\Gamma(n_{(q+1)k}+\alpha_k)}{\Gamma(n_{q+1}+\sum_{k=1}^{C}\alpha_k)}\right], 1\right\}, \quad (4.4.8)\end{aligned}$$

pour le changement de la variable explicative et de la règle de séparation.

L'algorithme des chaînes Monte Carlo markovienne cherche au travers de l'espace des arbres de classification pour trouver l'arbre désiré. Par construction, l'algorithme passe la majeure partie de ses itérations dans les régions avec la probabilité, $\Pr(\vec{T}|\vec{x},\vec{y})$, maximale. Selon Chipman *et al.* (2001), l'algorithme se stabilise rapidement dans un maximum et demeure dans cette région pendant un nombre d'itérations élevés. Nous savons aussi que l'algorithme est convergent et que celui-ci parcourra, éventuellement, l'espace des arbres de classification. Par

contre, à cause d'un temps limité et de la complexité de l'espace des arbres de classification, il faut augmenter la vitesse de convergence de l'algorithme.

Alors, Chipman *et al.* (2001) propose pour diminuer le temps de convergence de redémarrer l'algorithme plusieurs fois. Ainsi, en redémarrant l'algorithme au premier noeud, nous forçons celui-ci à effectuer un saut important dans sa probabilité *a posteriori* et à aller se promener, possiblement, dans une région différente. Une façon simple de déterminer le nombre d'itérations nécessaires avant un redémarrage est l'observation de la vraisemblance. Si la vraisemblance n'augmente plus d'une itération à l'autre, augmenter le nombre d'itérations peut être moins profitable que de redémarrer l'algorithme.

Pour compléter la méthodologie de la construction d'un arbre de classification sous un angle bayésien, il faut assigner une classe à chaque noeud terminal.

4.5. L'ASSIGNATION DES CLASSES AU NOEUDS TERMINAUX

La méthode utilisée la plus couramment consiste à déclarer la classe du noeud terminal comme étant celle où les observations se retrouvent en majorité. Par exemple, si nous avons dans un noeud terminal 5 cas atteints et 26 cas non atteints, nous déclarons le noeud comme non atteint. Donc, nous commettons une erreur pour les 5 cas atteints. Cette méthode marche bien quand, dans l'échantillon de départ, nous avons un nombre équivalent d'observations dans chacune des classes. Par contre, dans des situations plus extrêmes, où une classe peut représenter seulement 5 % des observations, il peut être avantageux d'attribuer d'une façon différente les classes aux noeuds. Nous proposons la méthode suivante (dans le cas où la variable dépendante ne prend que deux valeurs différentes, $C = 2$) :

$$I(y|\nu) = \begin{cases} 1, & \text{si } \Pr(y_i = 1|\nu) \geqslant \Pr(y_i = 2|\nu), \\ 2, & \text{si } \Pr(y_i = 1|\nu) < \Pr(y_i = 2|\nu), \end{cases} \tag{4.5.1}$$

où ν est la position du noeud terminal et $P(y_i = 1|\nu)$ est le nombre de cas de la classe 1 dans ce noeud divisé par le nombre de cas de la classe 1 présents dans

l'échantillon. Bref, la fonction indicatrice (4.5.1) assigne la valeur 1 lorsque le noeud terminal appartient à la classe 1 et la valeur 2 lorsque le noeud terminal appartient à la classe 2.

4.6. Le meilleur arbre

Une façon simple de choisir le meilleur arbre est de sélectionner celui qui possède la probabilité *a posteriori* maximale. Par contre, dans le cas de densités *a priori* qui ne tiennent pas compte des effets dilutions, cette méthode peut être problématique. Alors, si nous choisissons des densités *a priori* avec un effet dilution, nous utilisons seulement la vraisemblance $\Pr(\vec{y}|\vec{x},\vec{T})$ pour choisir le meilleur arbre, car celle-ci est indépendante de cet effet. De plus, en prenant l'arbre à vraisemblance maximale, nous pouvons maintenant choisir les densités *a priori* sans tenir compte de l'effet dilution.

En conclusion, dans ce chapitre, nous énonçons une méthode pour construire un arbre de classification en utilisant une approche bayésienne. Nous avons développé une fonction de vraisemblance explicite pour notre échantillon, $\Pr(\vec{y}|\vec{x},\vec{T})$, et nous avons défini toutes les fonctions *a priori* nécessaires au calcul de $\Pr(\vec{T})$. Par contre, vu que la probabilité $\Pr(\vec{T})$ ne se calcule pas explicitement, nous énonçons un algorithme utilisant les chaînes de Monte Carlo markoviennes à sauts réversibles pour optimiser la probabilité d'obtenir un arbre de classification, $\Pr(\vec{T})$. Ce choix vient du fait que nous avons un changement de dimension par la naissance ou la mort d'un noeud de l'arbre et que nous ne pouvons pas appliquer l'algorithme de Metropolis-Hastings directement. Bref, cette méthode consiste à générer des arbres de classification selon la probabilité *a posteriori* $\Pr(\vec{T}|\vec{x},\vec{y})$ et à conserver, si nous avons un effet de dilution dans les lois *a priori* choisies, l'arbre avec la probabilité de vraisemblance maximale, $\Pr(\vec{y}|\vec{x},\vec{T})$.

Dans le chapitre suivant, nous présentons les résultats des différentes méthodes, énoncées dans les chapitres 1, 3 et 4, appliqués aux données de la Banque Nationale.

Chapitre 5

RÉSULTATS

Dans les chapitres précédents, nous avons présenté plusieurs méthodes dans le but d'obtenir une performance de prévision adéquate sur le retard du remboursement d'un prêt investissement et ce, malgré l'absence d'un historique important. Nous avons donc appliqué ces différentes méthodes sur l'échantillon des prêts investissements décrit au chapitre 2 et nous avons mesuré leur performance.

Dans ce chapitre, nous commençons par construire une forêt d'arbres et nous énonçons une méthode de pondération des arbres de la forêt pour obtenir une seule probabilité de retard à chacun des clients. Ensuite, nous présentons l'arbre consensus qui nous permet de réduire la complexité de la forêt. De plus, pour comparer les différentes méthodes entre elles, nous énonçons deux mesures d'association. Finalement, nous présentons la performance de chacune des méthodes de prévision et nous concluons sur le modèle à conserver pour les prêts investissements.

5.1. Forêt d'arbres de classification bayésiens

Une forêt d'arbres est composée de plusieurs arbres où chacun attribue aux clients une probabilité d'être en retard sur le remboursement de son prêt investissement. Par conséquent, nous avons besoin, pour construire une forêt, d'un jeu de données différent d'un arbre à l'autre et d'une méthode pour pondérer les différentes probabilités. Dans ce mémoire, nous présentons un algorithme modifié de

la méthode dite « bagging » ou par vote qui a été introduite par Breiman (1996).

Tout d'abord, voici l'algorithme pour obtenir plusieurs échantillons :

(1) Construire une série U_i de N nombres provenant de la suite $\{1, \ldots, N\}$ avec remise, où N est le nombre d'observations de l'échantillon d'apprentissage.

(2) Ne conserver de l'échantillon d'apprentissage que les observations correspondant au numéro de la série U_i.

(3) Répéter les étapes (1) et (2) pour obtenir l'ensemble des séries $U = \{U_1, \ldots, U_D\}$, où D est le nombre d'arbres désiré.

Pour le choix du nombre d'arbres D à utiliser dans la forêt, Chipman *et al.* (2005) suggèrent de multiplier le nombre de variables importantes par 5, car utiliser plus de 5 arbres par variables importantes peut décroître le temps d'exécution pour une précision supplémentaire négligeable. Par exemple, si nous avons 10 variables importantes sur 200, nous allons générer 50 échantillons de l'algorithme précédent ($D = 50$). De plus, comme autre choix du nombre d'arbres, nous pourrions considérer seulement un nombre important d'arbres, par exemple $D = 1000$.

Ensuite, nous générons notre série d'arbres à partir des échantillons de U et de l'arbre de classification bayésien énoncé dans le chapitre 4. Alors, avec cette forêt, nous passons au travers de tous les arbres, notre échantillon d'apprentissage et notre échantillon d'évaluation. Ainsi, nous obtenons à chacune des observations une série de probabilités d'être en retard de longueur D.

En dernier lieu, nous exécutons une moyenne sur la différence des probabilités, c'est-à-dire :

$$\xi(y_i|\nu) = \frac{1}{D}\sum_{j=1}^{D}\left(\Pr(y_{ij} = 1|\nu) - \Pr(y_{ij} = 2|\nu)\right),$$

où $j \in \{1, \ldots, D\}$ représente l'arbre j de la forêt et y et ν sont définis dans la section 4.5. Nous utilisons la moyenne plutôt que la somme pour la fonction $\xi(y_i|\nu)$ afin que celle-ci soit comprise entre -1 et 1 et qu'elle ne dépende pas du nombre D. Ainsi, nous obtenons la prédiction suivante pour chacune des observations :

$$\varphi(y_i|\nu) = \begin{cases} 1, & \text{si } \xi(y_i|\nu) \geqslant 0, \\ 2, & \text{si } \xi(y_i|\nu) < 0. \end{cases}$$

Un inconvénient de la méthode dite « bagging » est que nous diminuons l'information de notre échantillon d'apprentissage en construisant plusieurs échantillons avec remise. Par contre, la forêt augmente significativement la performance de prévision comparativement à un seul arbre, (Chipman *et al.*, 2001).

5.2. ARBRE CONSENSUS

La forêt d'arbres augmente considérablement la complexité du modèle et de son implantation. Donc, nous proposons de calculer les fréquences des branches apparaissant dans la forêt et de reconstruire un arbre à partir des paires de noeuds les plus fréquentes.

L'algorithme est le suivant :

(1) Initialiser $i = 1$, $j = 1$ et $k = 1$.

(2) Calculer

$$\psi_{(ij|k)}(\vec{T}) = \sum_{l=1}^{D} I_{\left\{s_{il}^{pos} = s_{(\text{enfant de } i)l}^{pos}/2\right\} \cap \left\{s_{il}^{var} = s_{ij}^{var}\right\} \cap \left\{s_{(\text{enfant de } i)l}^{var} = s_{(\text{enfant de } i)j}^{var}\right\}},$$

$$\phi_{(ij|k)}(\vec{T}) = \sum_{l=1}^{D} I_{\left\{s_{il}^{pos} = (s_{(\text{enfant de } i)l}^{pos} - 1)/2\right\} \cap \left\{s_{il}^{var} = s_{ij}^{var}\right\} \cap \left\{s_{(\text{enfant de } i)l}^{var} = s_{(\text{enfant de } i)j}^{var}\right\}},$$

où $\psi(\cdot)$ calcule la fréquence pour chacune des paires de noeuds possibles où le noeud enfant est celui de gauche. La fonction $\phi(\cdot)$ calcule aussi la fréquence pour chacune des paires de noeuds, mais pour le noeud enfant de droite. De plus, s_{il}^{pos} et s_{il}^{var} sont respectivement la position et la variable de l'arbre l. L'indice i représente le noeud de la profondeur k, l'indice j

représente la paire de l'arbre j à comparer et $I_{\{\cdot\}}$ est une fonction indicatrice.

(3) Répéter l'étape (2) pour toutes les valeurs de $j \in \{2, \ldots, D\}$ sauf pour les paires $\left\{s_{ij}^{var} = s_{il}^{var}\right\} \cap \left\{s_{(\text{enfant de } i)l}^{var} = s_{(\text{enfant de } i)j}^{var}\right\}$ déjà comptées. Si le noeud est terminal, nous assignons par défaut la valeur 0 à $s_{l(\text{enfant de } i)}^{var}$ et $s_{(\text{enfant de } i)j}^{var}$.

(4) Répéter l'étape (2) et (3) pour toutes les profondeurs possibles de la forêt, $k = 1, \ldots, \max_{j \in 1,\ldots,D}(\rho_j)$ où ρ_j est la profondeur maximale de l'arbre j.

En suivant cet algorithme, nous obtenons deux séries de fréquences de paires par profondeur. Une série où le noeud enfant est à gauche et une autre série où le noeud enfant est à droite. Ensuite, pour obtenir notre arbre consensus, nous choisissons la paire avec un noeud enfant à droite la plus fréquente à la profondeur 1 et nous choisissons aussi la paire avec un noeud enfant à gauche la plus fréquente. Ensuite, nous découlons les autres profondeurs en choisissant toujours celle la plus fréquente de la série dépendamment de son noeud enfant (gauche/droite) et de sa profondeur. S'il y a égalité dans les fréquences et que le noeud terminal fait partie de ces paires, alors ce dernier a priorité. Les noeuds enfants nous indiquent les paires à choisir pour la profondeur correspondante.

Cet algorithme ne trouve que les valeurs s^{pos} et s^{var} pour l'arbre consensus. Il faut maintenant trouver les valeurs s^{regle} pour compléter notre arbre. Nous proposons de choisir la règle à chacun des noeuds qui maximise la vraisemblance de l'arbre. Finalement, nous pouvons construire plusieurs arbres consensus en choisissant la deuxième paire la plus fréquente comme noeud de départ ou en jouant avec les égalités de fréquences.

Nous présentons maintenant un exemple simple d'arbre consensus. Si nous avons une forêt à trois arbres, caractérisée par les vecteurs suivants :

(1) $\vec{T_1} = (1, 2, 8; 2, 1, 5; 3, 2, 4)$,

(2) $\vec{T_2} = (1, 1, 25; 3, 1, 5; 6, 2, 5)$,

(3) $\vec{T_3} = (1, 2, 4; 2, 1, 5)$.

Ces trois arbres sont dessinés aux figures 5.1, 5.2 et 5.3. Alors, en appliquant l'algorithme de l'arbre consensus, nous trouvons les fréquences des paires du tableau 5.1 et nous obtenons la première partie de l'arbre consensus $\vec{T} = (1, 2, x, 2, 1, x)$. Bref, il ne reste plus qu'à trouver les règles qui maximisent la vraisemblance et l'arbre consensus est, par exemple, $\vec{T} = (1, 2, 10, 2, 1, 6)$ est représenté à la figure 5.4.

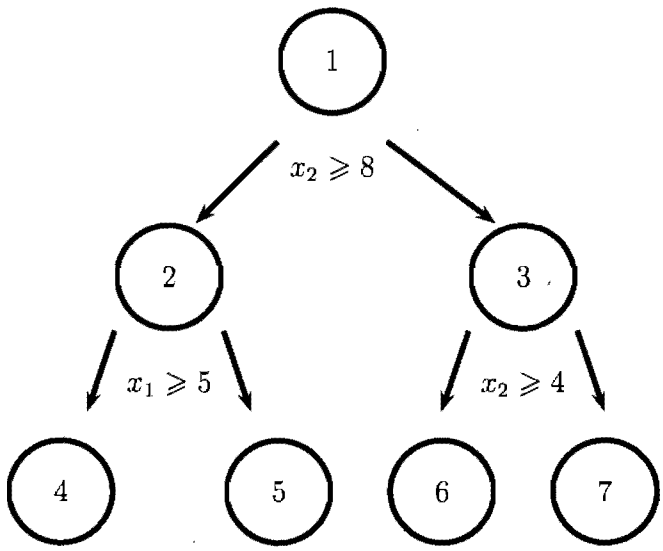


FIGURE. 5.1. Arbre de classification $\vec{T}_1$

5.3. MESURES D'ASSOCIATION

Pour évaluer la performance de prévision d'un modèle, nous avons besoin d'une mesure d'association. La mesure d'association nous permet de pondérer de différentes façons les cas mal classifiés. Nous présentons, dans ce mémoire, trois mesures d'association, soit la mesure ROC, où ROC est l'acronyme du terme anglais « Receiver Operating Curve », la statistique de Kuiper et le D de Somers.

5.3.1. La mesure ROC

La mesure ROC est une mesure d'association et elle sert à mesurer la performance de prévision d'un modèle. Une description de la mesure ROC peut être trouvée dans Song (1997). La mesure ROC calcule l'aire sous la courbe définie

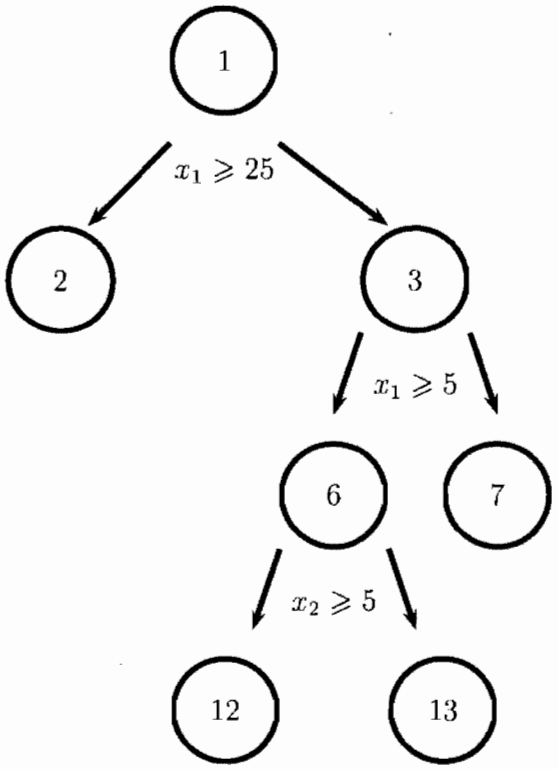


FIGURE. 5.2. Arbre de classification $\vec{T}_2$

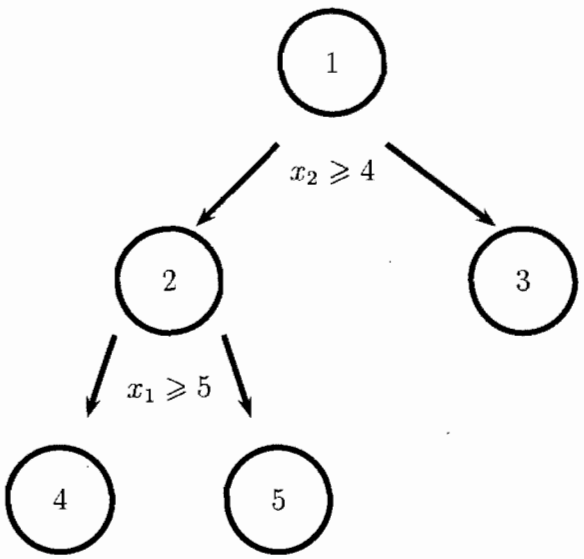


FIGURE. 5.3. Arbre de classification $\vec{T}_3$

par la sensibilité en ordonnée et un moins la spécificité en abscisse.

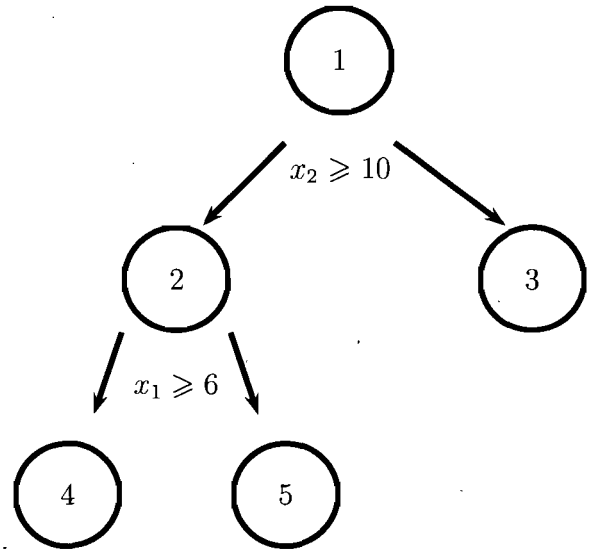


FIGURE. 5.4. Arbre consensus $\vec{T}$

TABLEAU. 5.1. Fréquences des paires de la forêt d'arbres.

Profondeur	Paire	Noeud pair/impair	Fréquence
1	(2,1)	pair	2
1	(1,0)	pair	1
1	(2,0)	impair	1
1	(1,1)	impair	1
1	(2,2)	impair	1
2	(1,0)	pair	2
2	(1,0)	impair	2
2	(2,0)	pair	1
2	(1,2)	pair	1
2	(1,0)	impair	1
2	(2,0)	impair	1
3	(2,0)	pair	1
3	(2,0)	impair	1

La sensibilité d'un test correspond à l'erreur de type 1, c'est-à-dire de rejeter hypothèse nulle sachant qu'elle est vraie et la spécificité d'un test correspond à l'erreur de type 2, c'est-à-dire de ne pas rejeter l'hypothèse nulle sachant qu'elle

est fausse.

Pour une prévision, nous allons plutôt définir la sensibilité et la spécificité par une proportion. Si nous voulons prédire une variable dépendante à deux classes (classe 0 absence, classe 1 présence), la sensibilité est la proportion des vrais absents sur les vrais absents plus les faux présents et la spécificité est la proportion des vrais présents sur les vrais présents plus les faux absents.

Bref, la sensibilité et la spécificité pour une prévision sont définies, de façon générale, comme suit :

$$\text{sensibilité} \quad : \quad F_1(z) = \frac{1}{n_1} \sum_{i \in C_1} I_{(\hat{\pi}_i \geq z)}, \tag{5.3.1}$$

$$\text{spécificité} \quad : \quad F_2(z) = \frac{1}{n_2} \sum_{i \in C_2} I_{(\hat{\pi}_i \geq z)}, \tag{5.3.2}$$

où $I_{\{\cdot\}}$ est une fonction indicatrice, n_1 est le nombre d'observations ayant une certaine condition, et n_2 est le nombre d'observations n'ayant pas cette condition. Le groupe de n_1 est dénoté par C_1, le groupe de n_2 est C_2 et $\hat{\pi}_i = \Pr(Y = 1|\vec{x})$ (section 1.4.1) est l'estimation de la probabilité de présence pour chacune des observations i. Pour chaque valeur de $z \in [0, 1]$, une sensibilité et une spécificité sont calculées. Pour une valeur de z fixée, nous obtenons un tableau similaire à 5.2 où la sensibilité et la spécificité sont calculées avec les équations suivantes :

$$\text{sensibilité} \quad : \quad F_1 = \frac{a}{a+b},$$

$$\text{spécificité} \quad : \quad F_2 = \frac{d}{c+d},$$

où a, b, c, d proviennent du tableau 5.2.

Donc, une sensibilité égale à 1 signifie que 100 % des observations ayant la condition absente ont été prédites correctement. Une spécificité égale 1 signifie que 100 % des observations ayant la condition présente ont été prédites correctement.

TABLEAU. 5.2. Tableau de prédiction permettant le calcul de la sensibilité et de la spécificité pour une valeur de z fixée.

	Prédiction du modèle		Total
	Absent	Présent	
Vraie valeur - Absent	a	b	$a+b$
Vraie valeur - Présent	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Maintenant que la sensibilité et la spécificité sont définies, nous pouvons calculer la mesure ROC par la formule suivante :

$$\text{ROC} = \left(n_c + 0{,}5\left(n_p - n_c - n_d\right)\right)/n_p, \tag{5.3.3}$$

où n_c est le nombre de paires concordantes, n_d est le nombre de paires discordantes et n_p est le nombre de paires. La mesure ROC est comprise entre 0 et 1. Ces paires se calculent en utilisant le tableau 5.2 comme ceci :

$$n_c = a \times d, \tag{5.3.4}$$

$$n_d = b \times c, \tag{5.3.5}$$

$$n_p = (a+b) \times (c+d). \tag{5.3.6}$$

En substituant les équations (5.3.4), (5.3.5) et (5.3.6) dans l'équation (5.3.3) et en manipulant l'équation résultante, nous obtenons la formule suivante :

$$\text{ROC} = \frac{1}{2}\left(\text{sensibilité} + \text{spécificité}\right) = \frac{1}{2}\left(\frac{a}{a+b} + \frac{d}{c+d}\right).$$

Alors, plus la mesure ROC est proche de 1, plus la classification est bonne et plus la mesure ROC est proche de 0, plus elle est mauvaise. Bref, pour une valeur de z fixée, la mesure ROC est une moyenne de la sensibilité et de la spécificité.

5.3.2. La statistique de Kuiper

Une description de la statistique de Kuiper peut être trouvée dans Kiefer (1959). La statistique de Kuiper se calcule seulement pour un échantillon ayant deux classes. Avant de définir cette statistique, nous avons besoin de la définition

d'une fonction de répartition empirique.

La fonction de répartition empirique pour un échantillon x_i où $i = 1, 2, \ldots, n$ et n est la taille de l'échantillon est notée comme suit :

$$F(x) = \frac{1}{n}(\text{nombre de } x_i \leqslant x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leqslant x),$$

où $I(\cdot)$ est une fonction indicatrice.

Maintenant que nous avons défini cette fonction, nous définissons l'échantillon de la classe 1 et de la classe 2 de la manière suivante :

$$\begin{aligned}(z_1, \ldots, z_{n_1}) &= (x_1, \ldots, x_{n_1}),\\ (z_{n_1+1}, \ldots, z_{n_1+n_2}) &= (y_1, \ldots, y_{n_2}),\end{aligned}$$

où n_1 et n_2 représentent les tailles échantillonales de la classe 1 et 2 respectivement. Ensuite, nous combinons les deux échantillons en un seul échantillon $(z_1, \ldots, z_{n_1+n_2})$. Finalement, nous pouvons calculer la statistique de Kuiper à l'aide de l'équation suivante :

$$K = \max_j(F_1(z_j) - F_2(z_j)) - \min_j(F_1(z_j) - F_2(z_j)),$$

où $j = 1, 2, \ldots, n_1+n_2$ et $F_1(x)$ et $F_2(x)$ sont les fonctions de répartition empirique de la classe 1 et 2 respectivement.

5.3.3. Le D de Somers

Une description du D de Somers peut être trouvée dans Somers (1962). Le D de Somers est une mesure d'association comme la mesure ROC et elle sert aussi à mesurer la performance de prévision d'un modèle. En utilisant les équations (5.3.1) et (5.3.2), nous pouvons calculer le D de Somers par l'équation :

$$D(z) = \text{sensibilité - (1-spécificité)} = F_1(z) - (1 - F_2(z)),$$

où $D(z) \in [-1, 1]$. De plus, nous pouvons aussi calculer le D de Somers en utilisant le tableau 5.2 et ce, pour une valeur de z fixée :

$$D = \frac{a}{a+b} - \left(1 - \frac{d}{c+d}\right) = 2 \times \text{ROC} - 1.$$

Alors, plus la valeur du D s'approche de 1, plus la qualité de classification s'améliore et plus D s'approche de -1, plus la qualité de la classification se déteriore. Bref, le D de Somers calcule la différence entre la sensibilité et spécificité et il n'est qu'une combinaison linéaire de la mesure ROC.

5.4. PERFORMANCES

Dans les chapitres précédents, nous avons construit sept méthodes différentes répertoriées ci-dessous avec leur section d'apparition :

(1) la régression logistique (1.4.1),

(2) la régression avec modèle « probit » (1.4.2),

(3) l'analyse discriminante linéaire (1.4.3),

(4) arbre de classification (3),

(5) arbre de classification bayésien (4),

(6) la forêt d'arbres de classification bayésiens (5.1),

(7) l'arbre consensus (5.2).

Nous avons appliqué les différents algorithmes explicités dans les chapitres 1, 3, 4 et 5 sur notre échantillon de données présenté dans le chapitre 2 et ce, dans le but de comparer les sept différentes méthodes entre elles. Pour chacune de ces méthodes, nous avons calculé la mesure ROC, la statistique de Kuiper et le D de Somers afin de mesurer la qualité de prévision de ces modèles.

Pour les méthodes (5), (6) et (7), nous avons appliqué une loi de densité *a priori* différente pour la variable contenant l'information de la valeur du fonds en garantie que celle énoncée dans la section 4.3.4.2. Ce nouveau choix de densité *a priori* est effectué, car nous avons de l'information supplémentaire sur cette variable que nous n'avons pas sur les autres. La construction de la densité *a priori* se

fait à partir de l'information que nous avons remarqué au chapitre 2, c'est-à-dire la relation significative entre la variable contenant la valeur du fonds et le retard sur le remboursement d'un prêt personnel.

Ainsi, nous proposons comme densité *a priori* pour la variable contenant l'information sur la valeur du fonds des prêts investissements, la fonction suivante :

$$\Pr(s_i^{regle}|(s_i^{var} = \text{valeur du fonds})) \propto \max_{j\in\{1,\ldots,\kappa\}} \left(\left(D_j(s_i^{regle}|s_i^{var})\right), \epsilon\right)$$

où ϵ est une valeur choisie arbitrairement petite et $D_j(s_i^{regle}|s_i^{var})$ est le D de Somers. La valeur ϵ vient donner une probabilité non nulle aux règles de séparation possédant un D de Somers négatif. Pour le choix du ϵ, nous référons le lecteur à la section 4.3.4.2. Nous proposons cette densité puisqu'elle attribue une probabilité non nulle plus importante aux valeurs de séparation donnant naissance à des noeuds homogènes. Le graphique de cette densité est représenté à la figure 5.5.

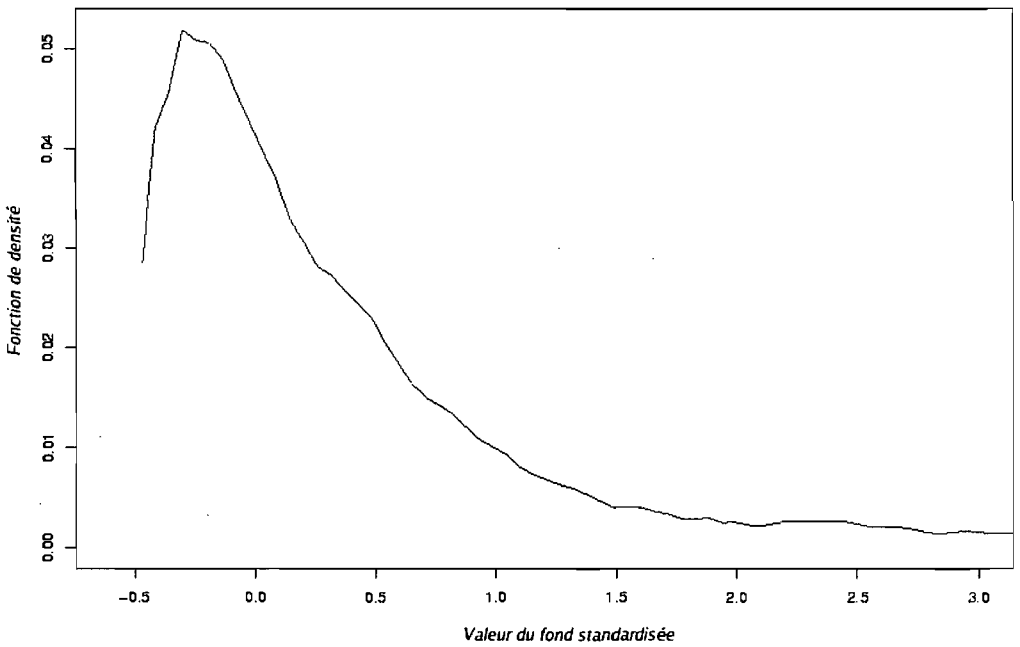


FIGURE. 5.5. Densité *a priori* pour la valeur du fonds standardisée

Pour la construction des arbres de classification bayésiens, nous avons testé sur notre échantillon les différentes fonctions *a priori* énoncées au chapitre 4. Ces fonctions sont rappelées ci-dessous et nous les nommons respectivement Wilcoxon-Somers, Position-Cauchy et Wilcoxon-Cauchy :

$$\begin{aligned}
\Pr_{\text{Wilcoxon-Somers}}(s_i^{var}) &= \frac{1-\text{valeur-p }(s_i^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))}, \\
\Pr_{\text{Wilcoxon-Somers}}(s_i^{regle}|s_i^{var}) &\propto \max\left(\left(D(s_i^{regle}|s_i^{var})\right),\epsilon\right), \\
\Pr_{\text{Position-Cauchy}}(s_i^{var}) &= \frac{\gamma_{s_i^{var}}}{\sum_{i=1}^{M}\gamma_{x_i}}, \\
\Pr_{\text{Position-Cauchy}}(s_i^{regle}|s_i^{var}) &= \frac{1}{\pi\gamma_i\left[1+\left(\frac{s_i^{regle}-\mu_i}{\gamma_i}\right)^2\right]}, \\
\Pr_{\text{Wilcoxon-Cauchy}}(s_i^{var}) &= \frac{1-\text{valeur-p }(s_i^{var})}{\sum_{j=1}^{M}(1-\text{valeur-p }(x_j))}, \\
\Pr_{\text{Wilcoxon-Cauchy}}(s_i^{regle}|s_i^{var}) &= \frac{1}{\pi\gamma_i\left[1+\left(\frac{s_i^{regle}-\mu_i}{\gamma_i}\right)^2\right]}.
\end{aligned}$$

Pour la forêt d'arbres de classification bayésiens, nous utilisons une forêt avec $D = 10$, soit 10 arbres, avec un RJ MCMC de 50 000 itérations, où une réinitialisation de l'algorithme a été faite après chaque tranche de 10 000 itérations. La lenteur du logiciel utilisé nous oblige à arrêter à 10 arbres, mais il serait possible d'augmenter ce nombre considérablement en changant de langage de programmation.

Pour les trois premières méthodes, soit la régression logistique, la régression avec modèle «probit» et l'analyse discriminante, nous nous sommes servis des procédures déjà programmées dans le logiciel SAS version 9.1.3. Pour les quatre autres méthodes concernant les arbres de classification, nous avons programmé entièrement les algorithmes en utilisant le langage de programmation R project version 2.2.1. De plus, nous avons programmé une fonction dessinant un arbre de classification. Nous présentons les quatre schémas d'arbres obtenus :

(1) l'arbre de classification à l'annexe C,

(2) l'arbre de classification bayésien Wilcoxon-Somers à l'annexe D,

(3) Position-Cauchy à l'annexe E,

(4) Wilcoxon-Cauchy à l'annexe F.

Nous présentons aussi les quatre arbres consensus :

(1) l'arbre de classification à l'annexe G,

(2) Wilcoxon-Somers à l'annexe H,

(3) Position-Cauchy à l'annexe I,

(4) Wilcoxon-Cauchy à l'annexe J.

Il est à noter qu'un arbre peut contenir deux noeuds terminaux provenant du même noeud et ayant la même classe. Par contre, même si ces noeuds terminaux ont la même classe, ils n'ont pas les mêmes probabilités d'être en retard ou sans retard.

Aussi, les résultats pour les deux mesures d'assocation ROC et Kuiper sont présentés pour les sept méthodes dans le tableau 5.3. Le D de Somers n'est pas présenté dans ce tableau puisqu'il peut s'écrire en une combinaison linéaire de la mesure ROC et par conséquent, il n'ajoute pas d'informations suplémentaires.

5.4.1. Analyse des résultats pour l'échantillon d'apprentissage

Par le tableau 5.3, nous analysons les résultats des différentes méthodes avec l'échantillon d'apprentissage. Nous remarquons que l'arbre bayésien (Wilcoxon-Somers) possède la mesure ROC la plus élevée et que la forêt d'arbres de classification possède la mesure de Kuiper la plus élevée. Ce résultat est contre-intuitif, puisque nous nous attendons à ce que la forêt d'arbres bayésiens performe mieux pour les deux mesures qu'un seul arbre et qu'une forêt d'arbres non bayésiens. Cette intuition se justifie par le fait que nous augmentons grandement la complexité du modèle en utilisant plusieurs arbres au lieu d'un seul et que nous pondérons ces arbres par leur qualité de prévision.

Nous expliquons ces résultas inattendus par le fait que nos différentes forêts contiennent seulement 10 arbres. Malheureusement, dans le contexte aléatoire

dans lequel ces forêts ont été crées, nous aurions eu besoin d'environ 180 arbres par forêt. Ce chiffre provient de la méthode de Chipman *et al.* (2005) qui suggèrent 5 arbres fois le nombre de variables explicatives de notre modèle. Dans notre cas, nous avons 36 variables explicatives pour un grand total de 180 arbres. Nous avons arrêté à 10 arbres puisque le temps et l'efficacité du programme ne nous permettaient pas dans obtenir davantage.

Pour l'échantillon d'apprentissage, voici les différentes conclusions :

(1) l'arbre bayésien performe mieux que l'analyse discriminante.

(2) L'approche bayésienne donne de meilleurs résultats que l'approche classique (l'arbre de classification, la forêt d'arbres et l'arbre consensus versus l'arbre bayésien, la forêt d'arbres bayésiens et l'arbre consensus bayésien).

(3) Si nous augmentons le nombre d'arbres dans la forêt, nous pouvons espérer augmenter les mesures d'association pour l'échantillon d'apprentissage.

Les résultats sur l'échantillon d'apprentissage sont moins importants, car il se peut que la qualité de prévision ne soit que superficielle à cause d'une trop grande complexité du modèle. Il faut donc analyser les résultats pour l'échantillon d'évaluation.

5.4.2. Analyse des résultats pour l'échantillon d'évaluation

Par le tableau 5.3, nous remarquons que la forêt d'arbres bayésiens (Wilcoxon-Cauchy) possède la statistique de Kuiper légèrement plus élevées que les autres méthodes. Par contre, la régression logistique possède une mesure de ROC plus élevée que les autres méthodes. Nous expliquons ce résultat par le fait que la régression logistique semble plus appropriée à la prédiction en utilisant des variables explicatives ordinales et que l'arbre de classification semble plus approprié pour des variables explicatives nominales (Joos, Vanhoof, Sierens et Ooghe, 1998).

Pour l'échantillon d'évaluation, voici les différentes conclusions :

TABLEAU. 5.3. Tableau des mesures d'association pour les différentes méthodes de prévision.

	Apprentissage		Évaluation	
Méthode	ROC	Kuiper	ROC	Kuiper
Régression logistique	0,664	0,296	0,653	0,319
Régression modèle «probit»	0,703	0,301	0,648	0,308
Analyse discriminante	0,592	0,276	0,558	0,287
Arbre de classification	0,585	0,170	0,554	0,108
Forêt arbres classification	0,558	0,472	0,544	0,305
Arbre consensus	0,593	0,185	0,553	0,105
Arbre bayésien (Wilcoxon-Somers)	0,713	0,426	0,622	0,245
Arbre bayésien (Position-Cauchy)	0,599	0,199	0,562	0,126
Arbre bayésien (Wilcoxon-Cauchy)	0,707	0,414	0,607	0,258
Forêt arbres bayésiens (Wilcoxon-Somers)	0,709	0,417	0,636	0,271
Forêt arbres bayésiens (Position-Cauchy)	0,657	0,382	0,621	0,253
Forêt arbres bayésiens (Wilcoxon-Cauchy)	0,697	0,412	0,644	0,330
Arbre consensus (Wilcoxon-Somers)	0,696	0,392	0,610	0,220
Arbre consensus (Position-Cauchy)	0,662	0,326	0,623	0,245
Arbre consensus (Wilcoxon-Cauchy)	0,629	0,369	0,558	0,293

(1) La forêt d'arbres bayésiens (Wilcoxon-Cauchy) et la régression logistique semblent donner des résultats similaires pour notre échantillon.

(2) L'approche bayésienne donne de meilleurs résultats que l'approche classique (l'arbre de classification, la forêt d'arbres, l'arbre consensus versus l'arbre bayésien, la forêt d'arbres bayésiens et l'arbre consensus bayésien).

(3) Si nous augmentons le nombre d'arbres dans la forêt, nous pouvons espérer augmenter les mesures d'association pour l'échantillon d'évaluation.

5.5. Choix du modèle

Après analyse des résultats, aucun modèle semble donner des mesures d'association systématiquement plus élevées que les autres. Par contre, vu la restriction importante sur le nombre d'arbres de nos différentes forêts, nous pouvons croire que les mesures d'association de ceux-ci pourraient augmenter de façon significative et, par le fait même, devenir systématiquement les plus élevées.

En conclusion, dans ce chapitre, nous avons présenté une forêt d'arbres de classification bayésiens construite à partir de la méthode «bagging» ou par vote. Nous avons énoncé un algorithme permettant de regrouper l'information d'une forêt d'arbres à un arbre consensus. Nous avons présenté trois mesures d'association, la mesure ROC, la statistique Kuiper et le D de Somers. Finalement, à l'aide de ces trois mesures d'association, nous avons comparé les résultats des septs méthodes. Nous avons trouvé que la forêt d'arbres bayésiens (Wilcoxon-Cauchy) est légèrement plus performante au niveau de la mesure de Kuiper de l'échantillon d'évaluation pour prédire le retard sur le remboursement d'un prêt investissement et qu'il faudrait augmenter le nombre d'arbres dans la forêt pour obtenir une meilleure mesure ROC.

Chapitre 6

CONCLUSION

Par ce mémoire, nous avons tenté de répondre aux deux objectifs fixés originalement. Ces objectifs consistaient en l'analyse de l'impact du comportement du fonds en garantie sur la délinquance du remboursement d'un prêt investissement et en la détermination de la probabilité de délinquance dans les 4 prochains mois pour chacun des clients.

Tout d'abord, nous avons récolté toutes les informations disponibles sur les clients ayant un prêt investissement à la Banque Nationale. Ensuite, nous avons regroupé ces informations dans une seule base de données et nous les avons analysées. Nous avons choisi de conserver 36 variables sur les 220 disponibles à cause de leur corrélation et de leur potentiel explicatif. De plus, nous avons imputé nos données, car nous avions un nombre important de valeurs manquantes et nous voulions conserver le maximum d'informations disponibles. Pour ce faire, nous avons comparé deux méthodes d'imputation et nous avons conservé la méthode avec le taux d'erreur le plus faible, soit l'imputation multiple.

En deuxième lieu, pour analyser l'impact du comportement du fonds en garantie, nous avons regardé la variation de la valeur du fonds et de son rendement dans le temps par rapport au comportement de la délinquance. Par des tests t pour échantillons indépendants, nous avons alors conclu qu'il ne semblait pas exister de lien entre la valeur du fonds en garantie ou le rendement pour les prêts investissements. Par contre, en regardant un prêt similaire, soit les prêts personnels,

nous avons trouvé une relation d'importance moyenne entre la valeur des fonds et la délinquance. Nous avons aussi trouvé pour les prêts personnels, en utilisant une régression logistique, que les fonds inférieurs à 1 000 ont approximativement une probabilité 2 fois plus élevée d'être en retard que ceux supérieurs à 150 000.

En troisième lieu, pour déterminer la probabilité de délinquance, nous avons testé sept modèles prédictifs différents. Nous avons, tout d'abord, présenté trois modèles linéaires fréquemment utilisés, soit la régression logistique de Hosmer, Lemeshow (2000), la régression avec modèle « probit » de Finney (1971) et l'analyse discriminante linéaire de Rao (1973). Par la suite, nous avons présenté en détails l'arbre de classification de Breiman *et al.* (1984) qui se trouve à être une méthode robuste et rapide tout en étant un modèle « boîte blanche ». Finalement, nous avons décidé d'élaborer l'arbre de classification en implantant l'inférence bayésienne pour créer les trois derniers modèles, soit l'arbre de classification bayésien de Schetinin (2004), Denison *et al.* (2002) et Chipman *et al.* (2001), la forêt d'arbres bayésiens de Breiman (1996) et l'arbre consensus.

Nous avons comparé les résultats de ces différents modèles prédictifs en utilisant les mesures d'association de ROC et de Kuiper. Ainsi, nous avons pu conclure que la forêt d'arbres bayésiens de Brieman (1996), construit avec les informations *a priori* disponibles et un nombre important d'arbres, semblait être la méthode la plus performante pour prédire le retard sur le remboursement d'un prêt investissement.

De plus, nous avons analysé le comportement des retards sur le remboursement d'un prêt puisque nous n'avons pas d'information disponible sur le défaut. Donc, il serait intéressant de mener une analyse similaire sur les défauts lorsque l'information sera disponible, car, d'un point de vue bancaire, la prédiction du défaut est plus importante que le retard.

Aussi, il serait intéressant de faire une étude de simulation sur les sept différents modèles de prévision dans le but de déterminer si la performance de la forêt d'arbres bayésiens est systématiquement meilleure que les performances des autres modèles. Finalement, d'autres modèles pourraient être envisagés pour prédire le retard ou le défaut, comme les réseaux de neurones et la méthode dite « Super Vector Machine (SVM) ».

BIBLIOGRAPHIE

[1] BISHOP Y.M.M., FIENBERG S.D., ET HOLLAND P.W.(1975), *Discret Multivariate Analysis :Theory and Pratice*, Cambridge, MA :The MIT Press.

[2] BREIMAN L.(1996), Bagging predictors *Machine Learning*, **24**(2), 123-140.

[3] BREIMAN L., FREIDMAN J.H., OLSHEN R.A. ET STONE C.J.(1984), *Classification and Regression Tree*, Wadsworth.

[4] CAMPBELL D.(1984), The computation of Catalan numbers, *Mathematics Magazine*, **57**, 195-208.

[5] CHIPMAN H., GEORGE E. ET MCCULLOCK R. (2005), BART : Bayesian additive regression trees, *IMS Lecture Notes, Monograph Series.*

[6] CHIPMAN H., GEORGE E. ET MCCULLOCK R. (2001), Bayesian CART model search, *IMS Lecture Notes, Monograph Series*, Volume 38.

[7] CHIPMAN H., GEORGE E. ET MCCULLOCK R. (1998), The pratical implementation of bayesian model selection, *J. American Statistics*, **93**, pp. 935-960.

[8] COHEN J.,(1988), *Statistical power analysis for the behavioral sciences (2nd ed.)*, Hillsdale, NJ : Lawrence Earlbaum Associates.

[9] DEMPSTER A. P., LAIRD N. ET RUBIN D. B. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[10] DENISON D., HOLMES C., MALICK B. ET SMITH A.(2002), *Bayesian Methods for Nonlinear Classification and Regression*, Wiley.

[11] DENISON D., MALICK B. ET SMITH A.(1998), A bayesian CART algorithm, *Biometrika*, **85**, 363-377.

[12] FINNEY D.J.(1971), *Probit analysis*, Third Edition, London : Cambridge University Press.

[13] GOULDEN, C. H. (1956), *Methods of statistical analysis*, Deuxième édition, New York Wiley.

[14] GREEN P.J.(1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination *Biometrika*, **82**, 711-732.

[15] HASTINGS W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, pp. 97-109.

[16] HOSMER D.W JR. ET LEMESHOW S. (2000), *Applied Logistic Regression*, Second Edition, New York, John Wiley & Sons, Inc.

[17] HOWLAND P., JEON M., PARK H. (2003), Structure preserving dimension reduction for clustered test data based on the generalized singular value decomposition, *SIAM Journal on Matrix Analysis and Applications*, **25**(1), p. 165-179.

[18] JOOS P., VANHOOF K., SIERENS N., OOGHE H. (1998), Credit classification : a comparison of logic models and decision trees, *Applications of machine learning and data mining in finance*, p. 59-73.

[19] KIEFER J. (1959), K-sample analogues of the Kolmogorov-Smirnov and Cramer-von Mises tests, *Annals of Mathematical Statistics*, **30**, p. 420-447.

[20] LI K.H (1988), Imputation using Marko chains *J. of Statistical Computation and Simulation*, **30**, pp. 57-79.

[21] LITTLE R.J.A. ET RUBIN D.B. (1987), Statistical Analysis with Missing Data, New York : John Wiley.

[22] METROPOLIS N., ROSENBLUTH A. W., ROSENBLUTH M. N., TELLER A. H., TELLER E., (1953), Equations of stat calculations by fast computing machines *J. Chem. Phys.*, **21**, pp. 1087-91.

[23] PARZEN E.(1962), On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, **33**, 1065-1076.

[24] QUINLAN J.R.(1986), Induction of decision trees, *Machine Learning*, **1**, 81-106.

[25] RAGEL A.(1998), Treatment of missing values for association rules, *Groupe de Recherche en Informatique Image et Instrumentation de Caen*, Université de Caen.

[26] RAO C.R.(1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York :John Wiley & Sons, Inc.

[27] RICE J.A.(1995), *Mathematical Statistics and Data Analysis*, Duxbury Press, 2ième édition, Californie.

[28] ROSENBLATT M.(1956), Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, **27**, 832-837.

[29] RUBIN D.B.(1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley & sons.

[30] SCHAFER J.L.(1997), *Analysis of Incomplete Multivariate Data*, New York, Chapman and Hall.

[31] SCHETININ (2004), The bayesian decision tree technique with a sweeping strategy, *Rapport Technique*, Coputer Science and Mathematics, University of Exeter.

[32] SOMERS R. H.(1962), A new asymmetric measure of association for ordinal variables, *American Sociological Review*, **27**, 799-811.

[33] SONG H.(1997), Analysis of correlated ROC areas in diagnostic testing, *Biometrics*, **53**(1), 370-82.

[34] SPATH H.(1980), *Cluster Analysis Algorithms*, Chichester, England :Ellis Horwood.

[35] TANNER M., ET WONG W. (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528-550.

[36] TIERNEY L. (1994), Markov chains for exploring posterior distributions, *The Annals of Statistics*, **22**, 1701-62.

[37] TIMOFEEV R.(2005), *Classification and Regression Tree, Theory and Applications*, Chichester, England :Ellis Horwood.

[38] WEYGANDT J.J., KIESO D.E., KIMMEL P.D., TRENHOLM B. (2003), *Principes de Comptabilité*, St-Laurent, Qc., ERPI.

[39] WILCOXON, F. (1945), Individual comparaisons by ranking methods, *Biometrics*, **1**, 80-83.

Annexe A

CORRÉLATIONS

Variables	Valeur p	Corrélation avec le retard
Age du client	0,154	-0,012
Valeur prêt/valeur fonds	0,787	-0,002
Score risque actuel	0,003	-0,024
Score risque précédent	<0,001	-0,033
Volatilité portefeuille	0,695	0,003
Valeur marchande	0,269	0,009
Valeur prêt-valeur fonds	0,854	-0,002
Score à l'octroi	0,001	-0,038
Comportement octroi	0,905	-0,001
Variation prêt/fonds	0,902	-0,001
Retard 90+jrs octroi	0,003	0,024
Retard -30jrs octroi	<0,001	0,043
Retard -60jrs octroi	<0,001	0,030
Retard -90jrs octroi	<0,001	0,029
Revenu principal	0,751	0,0026
Paiements mensuels (rotatif)	0,481	0,006
Écart-type Rendement	0,071	0,015
Total engagement	0,726	-0,003

Variables	Valeur p	Corrélation avec le retard
Retard 1 an	<0,001	0,133
Retard temps	<0,001	-0,194
Retard fréquence	<0,001	0,194
Rendement du fonds	0,024	-0,018
novembre 2004	0,137	0,012
décembre 2004	<0,001	0,041
janvier 2005	<0,001	0,030
février 2005	<0,001	0,055
mars 2005	<0,001	0,060
avril 2005	<0,001	0,047
mai 2005	<0,001	0,045
juin 2005	<0,001	0,103
juillet 2005	<0,001	0,101
septembre 2005	<0,001	0,158
octobre 2005	<0,001	0,241
août 2005	<0,001	0,058

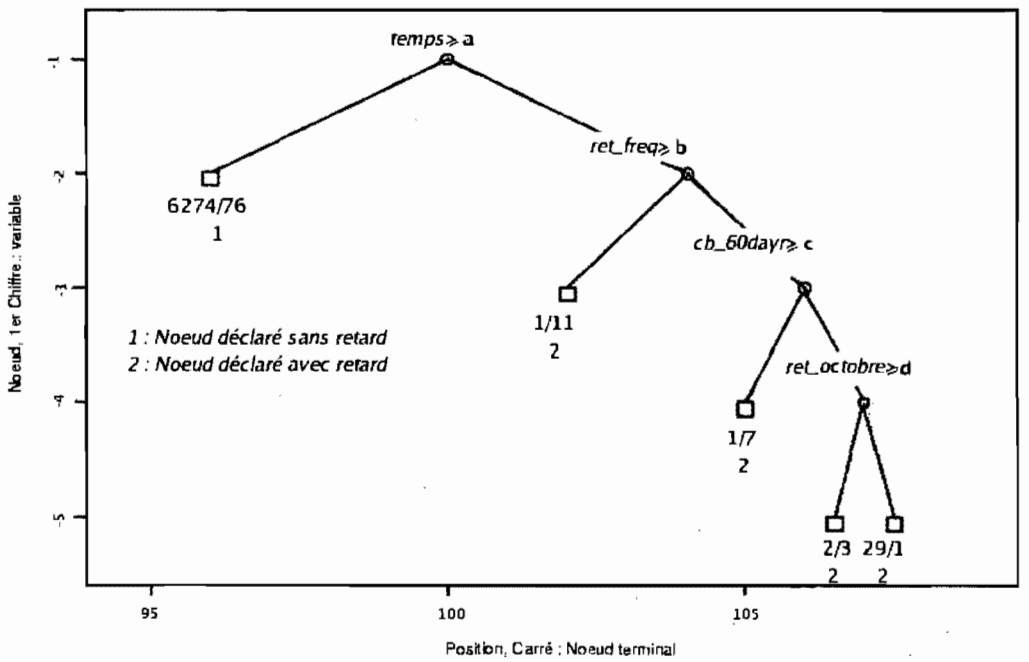
Annexe B

VALEURS MANQUANTES

Variables	Nb. valeurs manquantes	Fréquence relative
Age du client	44	0,294
Score risque actuel	922	6,140
Score risque précédent	1805	12,079
Score à l'octroi	6227	41,671
Comportement octroi	6227	41,671
Retard 90+jrs octroi	6227	41,671
Retard -30jrs octroi	6227	41,671
Retard -60jrs octroi	6227	41,671
Retard -90jrs octroi	6227	41,671
Revenu principal	6227	41,671
Paiements mensuels (rotatif)	6227	41,671
Total engagement	6227	41,671

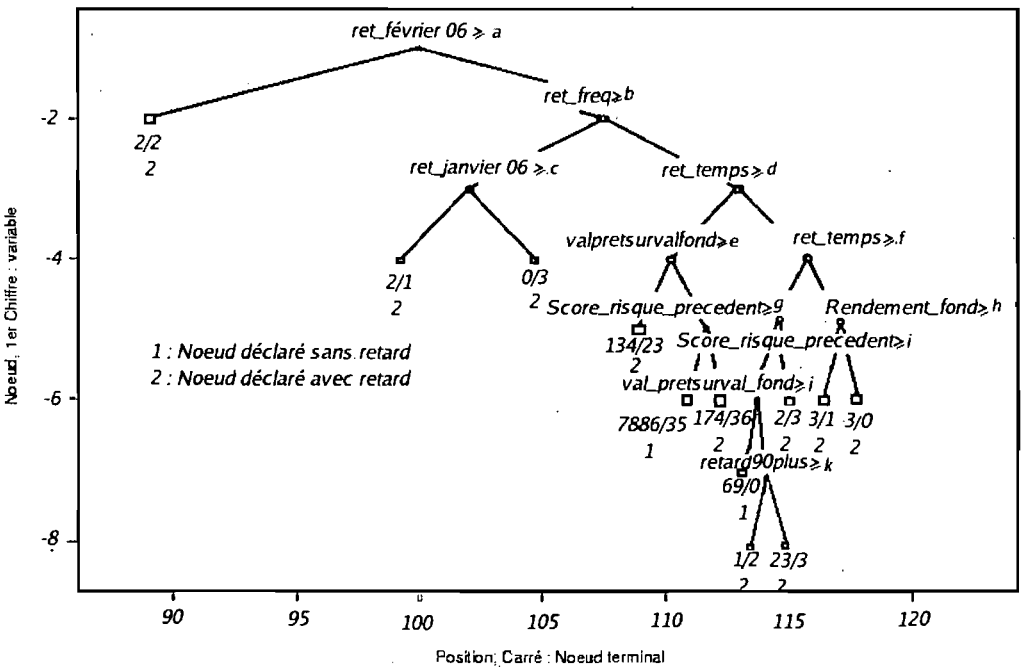
Annexe C

UN ARBRE - L'ARBRE DE CLASSIFICATION



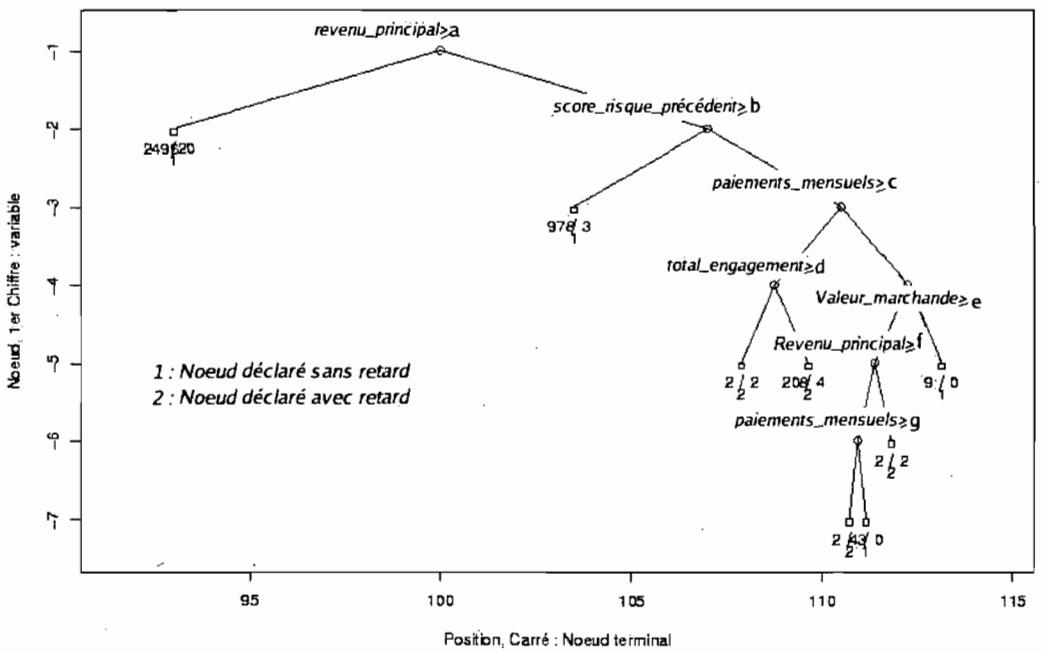
Annexe D

UN ARBRE - BAYÉSIEN WILCOXON-SOMERS



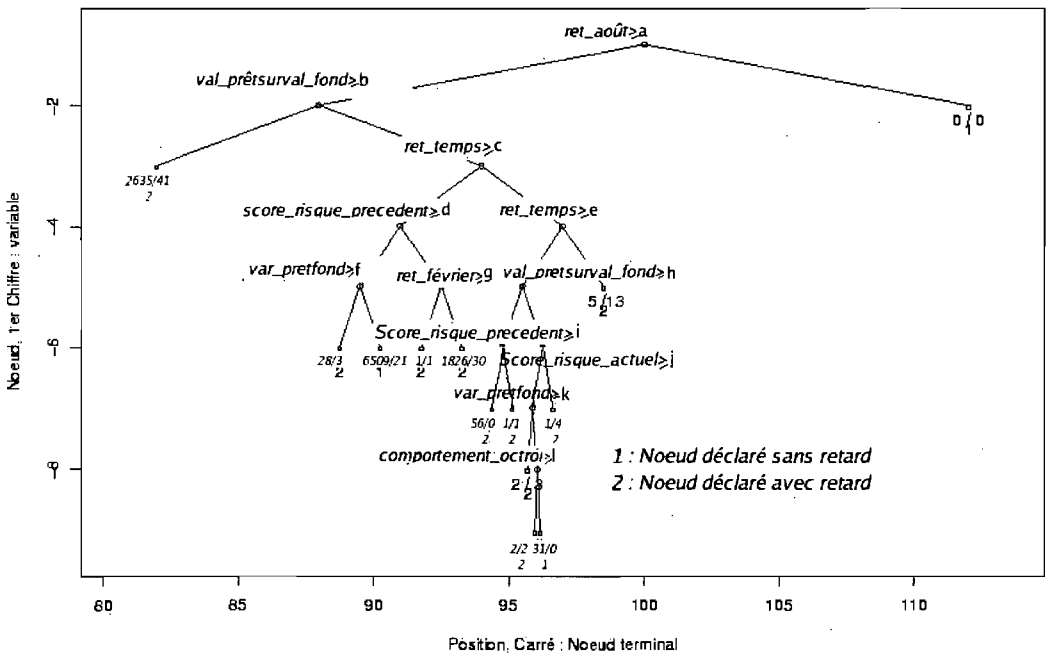
Annexe E

UN ARBRE - POSITION-CAUCHY



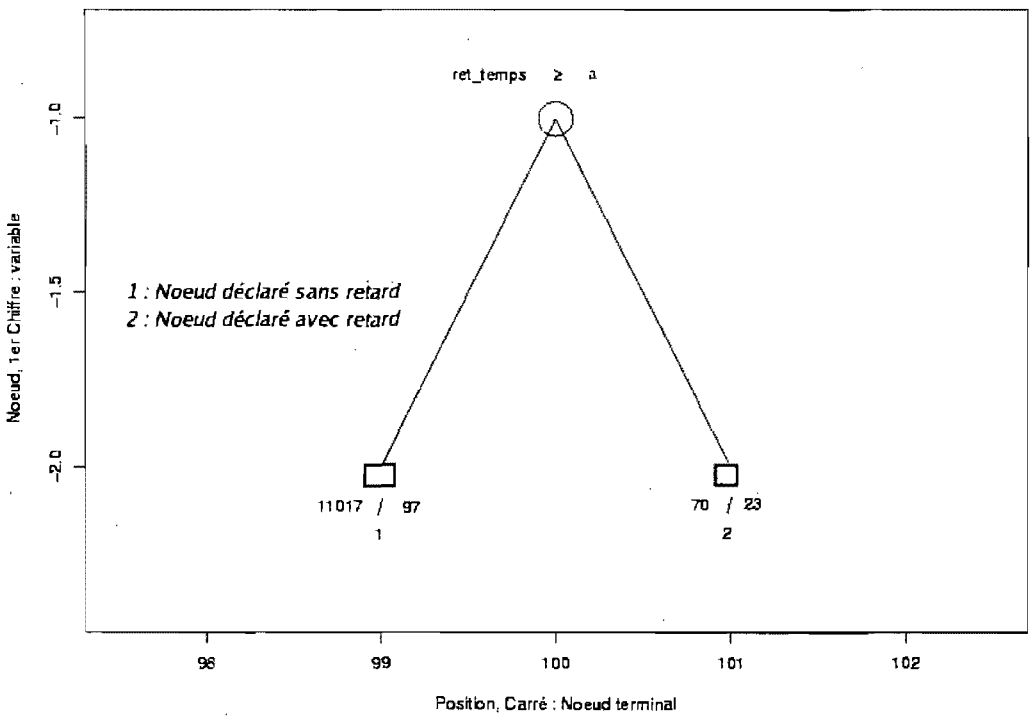
Annexe F

UN ARBRE - WILCOXON-CAUCHY



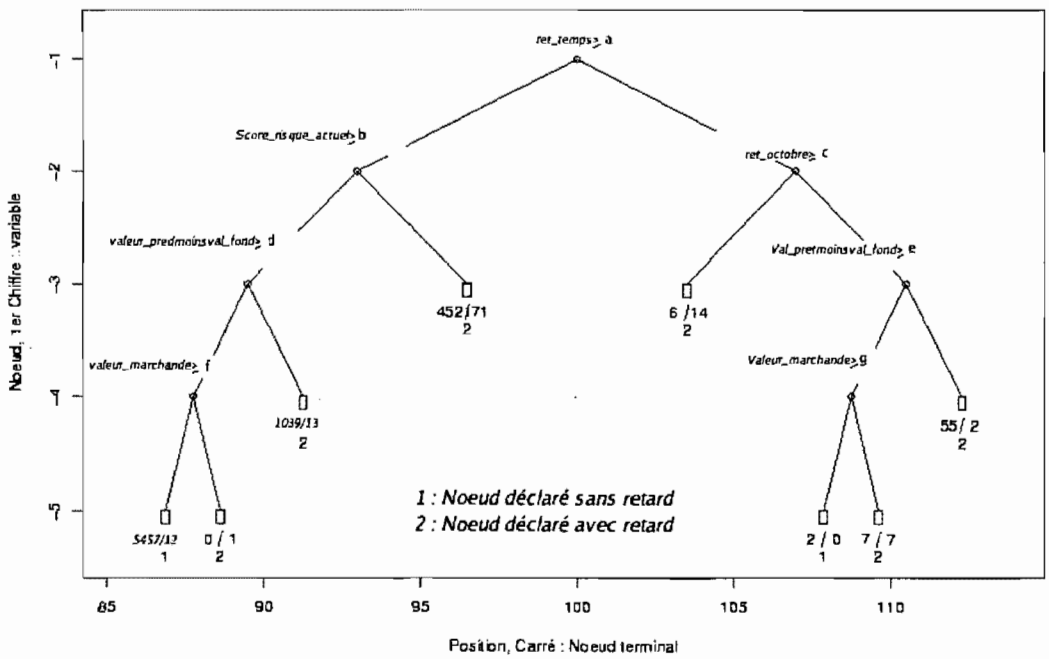
Annexe G

ARBRE CONSENSUS - L'ARBRE DE CLASSIFICATION



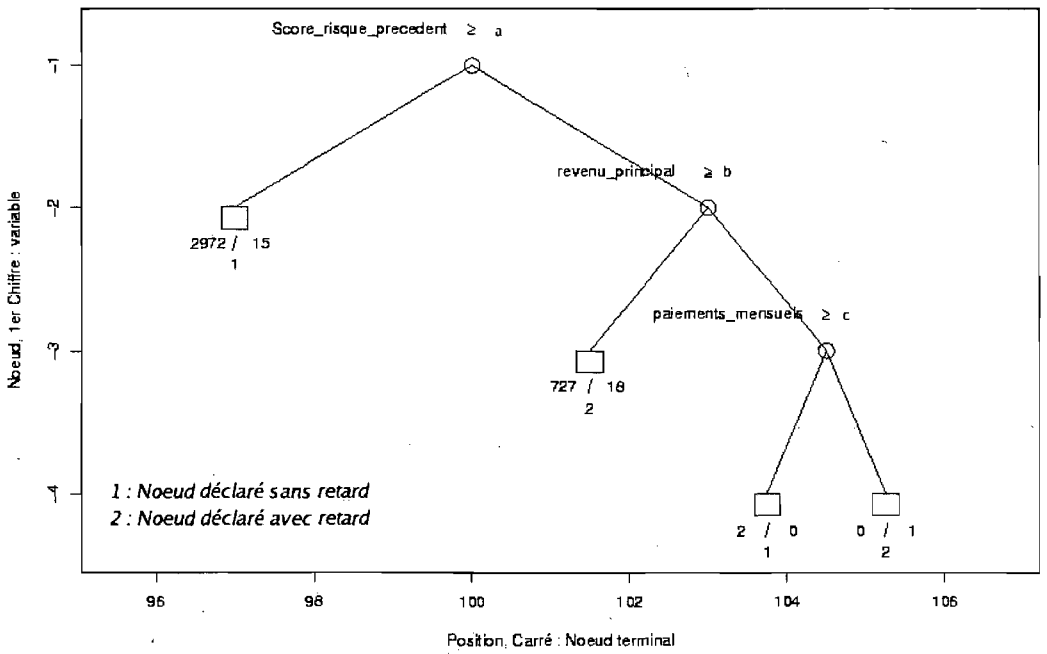
Annexe H

ARBRE CONSENSUS - BAYÉSIEN WILCOXON-SOMERS



Annexe I

ARBRE CONSENSUS - POSITION-CAUCHY



Annexe J

ARBRE CONSENSUS - WILCOXON-CAUCHY